Program Chairs' Report on Peer Review at ACL 2023. **Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, Naoaki Okazaki**. *Association for Computational Linguistics*, 2023, 33 pages.


```
    @inproceedings{2023-2023,
Author = {Program Chairs' Report on Peer Review at ACL 2023},
Title = {Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, Naoaki Okazaki},
Journal = {Association for Computational Linguistics},
Year = {2023},
Location = {Toronto},
Url = {http://umiacs.umd.edu/~jbg//docs/2023_acl_peer_review_report.pdf},
}
```


Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Program Chairs' Report on Peer Review at ACL 2023

Anna Rogers[♢] **Marzena Karpinska**[♡] **Jordan Boyd-Graber**[♠] **Naoaki Okazaki**[♣]

[♢]IT University of Copenhagen [♡]University of Massachusetts Amherst
[♠]University of Maryland [♣]Tokyo Institute of Technology

arog@itu.dk mkarpinska@cs.umass.edu
jbg@umiacs.umd.edu okazaki@c.titech.ac.jp

Abstract

We present a summary of the efforts to improve conference peer review that were implemented at ACL'23. This includes work with the goal of improving review quality, clearer workflow and decision support for the area chairs, as well as our efforts to improve paper-reviewer matching for various kinds of non-mainstream NLP work, and improve the overall incentives for all participants of the peer review process. We present analysis of the factors affecting peer review, identify the most problematic issues that the authors complained about, and provide suggestions for the future chairs. We hope that publishing such reports would (a) improve transparency in decision-making, (b) help the people new to the field to understand how the *ACL conferences work, (c) provide useful data for the future chairs and workshop organizers, and also academic work on peer review, and (d) provide useful context for the final program, as a source of information for meta-research on the structure and trajectory of the field of NLP.

1 Introduction

With the continued growth of our field and the rising number of conference submissions, peer review draws more and more attention from the community—as an application area (Hua et al., 2019; Anjum et al., 2019; Stelmakh et al., 2019, inter alia), in meta-research (Rogers and Augenstein, 2020; Church, 2020, inter alia), in initiatives to organize and release peer review data (Kang et al., 2018; Jecmen et al., 2022; Dycke et al., 2022, inter alia), and, of course, in the regular heated social media discussions during submission deadlines, review release dates, and acceptance notifications. It is unlikely that peer review will ever be perfect – it remains 'the least bad system' we have for ensuring the quality of scientific publications (Smith, 2010). Still, with each iteration we should learn a little more about what works better for organizing peer review at such scale, and in a community so diverse in expertise and experience.

As a step in that direction, ACL'23 makes its peer review report public and an official part of the conference proceedings, complementing the introduction and other administrative materials. The goal is to increase the visibility of the results of the conference process, as well as any incidental findings from conference organizations and the lessons learned the hard way that may be useful to the future chairs and workshop organizers. Such publications also provide extra incentives for the future program chairs to invest more effort in the analysis of their process, and they provide a useful background to the composition of the final program that may be useful for meta-science research (since they essentially document the selection process for that program). Last but not least, such publications will improve the transparency of the *ACL conference process, which may be useful to the researchers who are new to the field.

We present the core statistics per track (§2), analysis of resubmissions (§3) and core demographics (§4), our efforts for improving peer review quality (§5), improving decision support for the chairs (§6), out analysis of various factors contributing to review scores and final decisions (§7), ethics review and best paper selection (§8), and our efforts towards improving incentives for the authors, reviewers and chairs (§9). We conclude with overall recommendations for future conference organizers (§10). The materials we developed will be available at a dedicated repository[1].

The results presented here are based on the analysis of internal data of ACL'23, as well as exit surveys that we sent to the chairs, authors and reviewers. We received responses from 25 senior area chairs (SACs)

[1]https://github.com/acl-org/acl-2023-materials

Track	*Direct submissions*			*ARR submissions*		
	Submitted	Main	Findings	Submitted	Main	Findings
Computational Social Science and Cultural Analytics	113	22.12	19.47	10	90.00	10.00
Dialogue and Interactive Systems	269	24.54	15.24	19	21.05	42.11
Discourse and Pragmatics	52	21.15	34.62	1	100.00	0.00
Ethics and NLP	54	22.22	31.48	7	42.86	42.86
Generation	175	25.71	20.57	6	66.67	16.67
Information Extraction	279	25.45	16.13	33	24.24	36.36
Information Retrieval and Text Mining	94	14.89	21.28	9	44.44	0.00
Interpretability and Analysis of Models for NLP	189	24.34	28.04	20	35.00	55.00
Language Grounding to Vision, Robotics, and Beyond	147	24.49	21.77	5	40.00	40.00
Large Language Models	252	28.17	21.03	10	50.00	30.00
Linguistic Diversity	18	27.78	22.22	1	0.00	100.00
Linguistic Theories, Cog. Modeling & Psycholinguistics	38	23.68	23.68	8	50.00	37.50
Machine Learning for NLP	313	21.09	23.32	37	56.76	2.70
Machine Translation	198	25.25	18.18	7	0.00	57.14
Multilingualism and Cross-Lingual NLP	85	20.00	30.59	12	25.00	16.67
NLP Applications	354	22.88	19.77	25	52.00	8.00
Phonology, Morphology, and Word Segmentation	21	28.57	19.05	0		
Question Answering	197	18.78	18.78	22	45.45	18.18
Resources and Evaluation	213	28.17	19.72	23	56.52	0.00
Semantics: Lexical	54	25.93	25.93	3	66.67	33.33
Semantics: Sentence-level Semantics	81	27.16	11.11	9	22.22	22.22
Sentiment Analysis, Stylistic Analysis, Arg. Mining	107	17.76	30.84	10	30.00	0.00
Speech and Multimodality	72	27.78	36.11	7	57.14	14.29
Summarization	139	23.02	21.58	12	33.33	8.33
Syntax: Tagging, Chunking, and Parsing	69	23.19	21.74	5	20.00	20.00
Theme: Reality Check	110	26.36	30.91	1	100.00	0.00
Total	4559	20.73	18.36	305	42.30	20.98

Table 1: Number of submissions and acceptance rates per track for direct and ARR submissions to ACL'23.

(35.7% response rate), 134 area chairs (ACs) (30.5% response rate), 510 reviewers (11.4% response rate), and 556 authors (4.07% response rate of all authors[2]).

2 Tracks and Acceptance Statistics

ACL'23 had 26 tracks, most of which have also been offered at other recent NLP conferences. At the suggestion of EMNLP 2022 chairs, we kept their separation of "*Large Language Models*"[3] track from "*Machine Learning for NLP*" track. At community requests we added the following tracks: "*Linguistic Diversity*" and "*Multilingualism and Cross-lingual NLP*". Each track had at least two Senior Area Chairs (SACs), who then recruited area chairs (ACs) for that track. The full list of senior chairs per track is available at the conference website.[4]

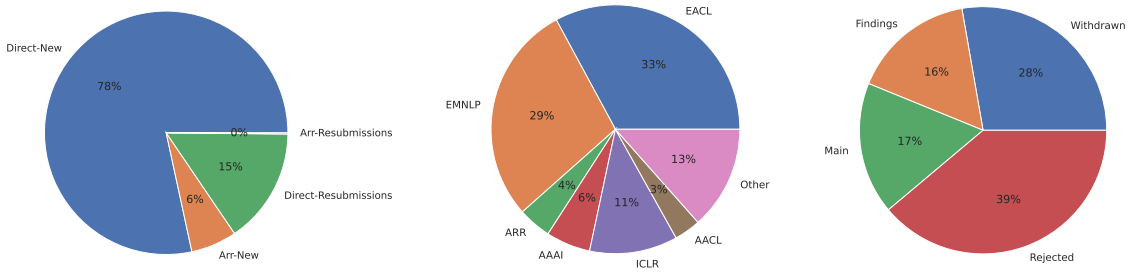
Internally, in the START system there were also two special tracks: "*Ethics review*" track (which handled the reviews of papers that were flagged for ethical issues), and "*Conflicts of interest*" (COI) track, which handled the papers with which the SACs of the relevant tracks had a COI.

ACL'23 implemented a hybrid process, in which it was possible to submit papers either directly to the START system (to be reviewed through ACL'23 internal peer review process to be described in this report), or commit it through ACL ROlling Review (ARR) with reviews already performed at ARR. Most submissions to ACL'23 were direct submissions (4559), and 305 more came through ACL Rolling Review (ARR). Table 1 shows acceptance for each type of submission and in each track.

[2]Assuming that in most cases at most one author per paper responded to the survey, the upper bound on the response rate for author feedback per paper would be 11.4% of all direct and ARR submissions that were reviewed. 37.9% of the authors who responded to the survey indicated that they disagreed with the outcome for their submission.

[3]The EMNLP original name was *Language Modeling and Analysis of Language Models.*. In our version it was simply *Large Language Models*, as they are the most frequent topic currently, but in retrospect the original version is preferable as it is more inclusive.

[4]https://2023.aclweb.org/committees/program/



(a) Submissions vs resubmissions (b) Prior venues of resubmissions (c) The fate of resubmitted papers

Figure 1: Resubmissions at ACL'23

ACL Rolling Review (ARR). Table 1 shows that in most tracks, ARR submissions had a much higher acceptance rate, sometimes twice higher. This is to be expected because ARR submissions self-select for high scores and positive reviews before committing to ACL.

Since in the hybrid process ARR submissions and direct submissions directly compete for acceptance, a question arises to what extent this is a fair competition. We asked that question to our SACs. 58.3% believe that this process is fair enough, 12.5% - that it is unfair to the direct submissions, and 29.6%—that it is unfair to the ARR submissions. Of 17 SACs who believed that this situation is unfair in some way, 23.5% suggested that they should have separate acceptance rate, 41.2%—that they should have a separate process and acceptance criteria, and 47.1%—that there should be some other solution (many comments pointing to the confusion, the apples-to-oranges comparisons of reviews performed with different evaluation, the less-than-ideal import of openreview data into START (browsing attachments takes more time). Many expressed a preference for a non-hybrid process.

As program chairs, our biggest challenge with ARR was that by design it provides reviews and meta-reviews, but the acceptance decisions are then made by our SACs—who generally do not provide extra feedback to either direct submissions or ARR submissions (nor can they be expected to: some tracks had over 300 papers per 3 SACs). For direct submissions, nobody expects SAC-level feedback. But to ARR authors, who likely self-selected for high scores and positive reviews, to be rejected without explanation is more frustrating, and we received a lot of angry emails demanding extra feedback (even though neither we nor ARR promised that). It seems that by design, a process where there are acceptance quotas, and decisions are fully decoupled from feedback, will necessarily leave the majority of authors rejected without explanation—and hence disappointed and unsure what they could do to improve their work (and we agree that this would indeed be frustrating to the authors).

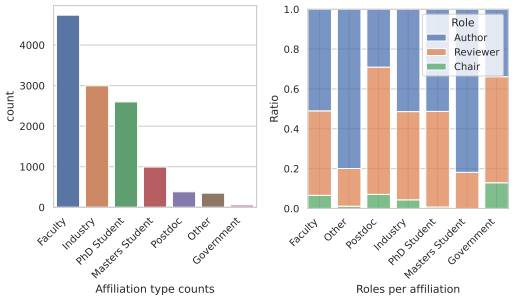
The above factors could transform into a bigger problem in the future. We only had 305 ARR submissions, but if a majority of our submissions came with high scores and positive reviews—this just would not be a useful signal anymore. The acceptance odds of direct submissions would decrease (as compared to a process where everyone starts at the same stage of peer review). The SAC-ing would become harder (since selecting among high-quality papers is less easy than among papers of varying quality), and the authors would be disappointed because many would be rejected with high scores and no idea what they could do differently.

3 Resubmissions

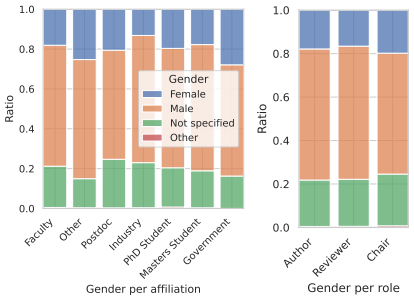
Among the 4559 direct submissions to ACL'23, 754 indicated that they were resubmissions (see fig. 1a). The biggest "donors" were EACL[5] (296), EMNLP (258), ICLR (103), AAAI (52), and ACL Rolling Review[6] (39). Although the selectivity of top-tier conferences means that the majority of papers are

[5]Because our submission deadline was shortly before EACL and ICLR notification deadlines, we made an exception to no-cross-submission policy and allowed their submissions to be also submitted to ACL. After their respective notifications many such papers withdrew from our pool, which explains the high withdrawal rate in Figure 1c.

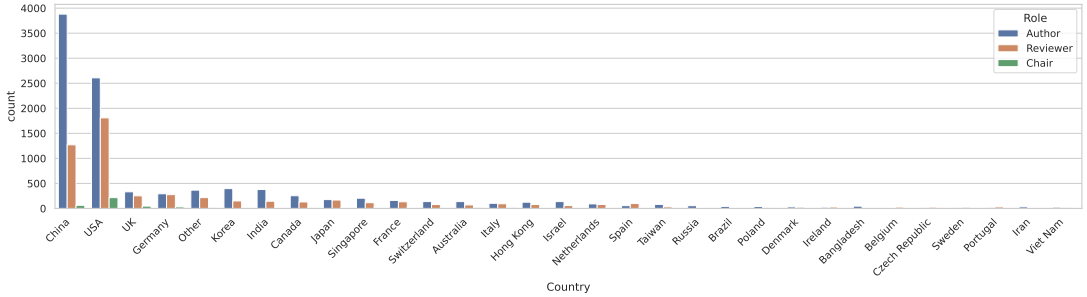
[6]There were 11 resubmissions from October 2022, 6 from September, and 1-3 from many other months of 2022.



(a) Affiliation types distribution



(b) Gender distribution



(c) Top 30 countries listed in the ACL author and reviewer profiles

Figure 2: Author and reviewer pool at ACL'23*

* All information is self-reported, not independently verified, and does not correspond to any specific definition of affiliation, gender, or country (e.g., some authors from Edinburgh may elect to list their country as "Scotland" rather than "UK".)

rejected, the bulk of the ACL'23 submissions are new, which means that at this point **the burden of re-reviewing is relatively low**. It is possible that this is due to the wider acceptance of Findings as a publication channel, as more *ACL conferences continue to offer this option.

Moreover, ACL'23 authors had the option to submit previous reviews as an attachment, but only 243 submissions used this option, which suggests that most resubmitters preferred to have a completely new set of reviewers. ARR allows that option within ARR, but the ARR submissions themselves did not have a high rate of revise-and-resubmit (only 8/305), as shown in fig. 1b.

Intuitively, one could expect that resubmissions have a higher chance of acceptance, since these are the papers that have received feedback and had a chance to revise. But fig. 1c suggests otherwise. See more analysis in §7.3.

4 Authors and Reviewers at ACL'23

We received a record 4864 submissions (4559 direct, 305 from ARR) from the total of 13,658 authors, reviewed by 4490 reviewers. This section reviews our recruitment process and the three demographic variables (country, affiliation type, and gender) to which we had access in the global START profiles of all participants of ACL peer review process.

Reviewer recruitment. We initially sent review invitations to the reviewer list which we had received from the organizers of previous conferences. We also required the authors of all submissions to nominate at least one experienced reviewer, whom we also sent invitations.

As we elicited reviewer data, we found that **for a quarter of our reviewers**[7] **there is no reliable Semantic Scholar publication history data that can be used for paper-reviewer matching**. For conferences that fully rely on automated paper–reviewer matching based on publication history, this factor obviously sets a bound on their possible performance. Often the author pages exist because Semantic Scholar automatically created them, but the authors did not claim them and did not clean them up, which

[7] Out of the reviewers who filled in our sign-up forms, only 75.4% confirmed that their Semantic Scholar profile is accurate and can actually be used to estimate their areas of interest and expertise. In addition to that, 8.9% reviewers listed in START did not specify their Semantic Scholar IDs in their profiles.

may result in the addition of publications by namesake authors (e.g. the automatically created profile for "Anna Rogers" originally had contributions from at least three researchers with that name.) This is particularly worrying because at this point many venues have used this information for paper-reviewer matching, and urged the NLP community to maintain their Semantic Scholar profiles. We also specifically reminded about this, but still a quarter of our sign-up pool stated that their publication history is not accurate. In addition to this problem, matching based on publication history has the issue with establishing expertise of different authors on on multi-author publications. Hence, we developed an alternative matching approach described in §5.2.

Affiliation types. Figure 2a presents the overall distribution of the affiliations of our authors and reviewers (as stated in START profiles). The biggest group of authors, reviewers, and chairs are academic faculty. The second biggest group (by absolute numbers) in all three categories is industry, which is relevant to the recent concerns about the influence of industry on academic NLP research (Abdalla et al., 2023). Furthermore, students form at least 26% of reviewer pool (Ph.D. 22.7%, M.Sc. 3.3%). This was also our experience as area chairs at other recent conferences, and it highlights the need to **continue the reviewer training efforts**.

Gender distribution. Based on the information in softconf profile, about 20% of ACL peer review participants in all roles did not answer the question about their gender (Figure 2b). For a part of this population this is likely a deliberate choice, but judging by how many other fields in the START profiles were not accurately filled in or updated, in many cases this likely signals simply the lack of desire to fill in forms, especially for the new authors who had to register in START last minute in order to make a submission. Considering only those profiles that responded to this question, we see a heavy imbalance for "male", in agreement with the reports on under-representation of women in Computer Science (Jaccheri et al., 2020; Pantic and Clarke-Midura, 2019), where a lot of NLP research is currently happening. This underscores the need to **continue the Diversity and Inclusion efforts**.

Top contributing countries. The analysis of the countries of all authors and reviewers suggests that the balance between reviewing and submitting papers is considerably off for many locations, and particularly China.[8] We believe that this is at least partly due to the fact that our recruitment efforts started with the pool of the previous conferences. That pool needs to be deliberately expanded by **more active and targeted reviewer recruitment efforts among Chinese institutions**.

Church (2020) estimates that at 20% acceptance rate the authors of published papers "owe" the community at least 15 reviews per each publication (3 for their own paper, and 4x3 for the papers that didn't get in). While some dis-balance between the author and reviewer list is to be expected (e.g., since many junior authors are not yet qualified to review, and many senior authors perform other organization roles)—we clearly need to decrease it in order to decrease the reviewer load. Our default quota was six papers[9] per reviewer, in line with most recent conferences. This is a significant workload, and it can hardly be expected to improve the quality of reviews. Moreover, the more reviewers are in the pool, the smaller the trade-off between optimizing for best matches or smaller workload per reviewer.

5 Efforts towards improving review quality

This section describes the following steps that ACL'23 proposed and implemented within its peer review process to improve review quality: review tutorials (§5.1), Area-Contribution-Language paper-reviewer matching (§5.2), flagging of review issues by the authors (§5.3). The efforts to improve the overall incentives are decribed in §9.2 and §9.3.

[8] In absolute numbers: 3881 authors vs 1271 reviewers for China (ratio 3.05, absolute difference 2610). For the US: 2608 authors, 1809 reviewers (ratio 1.4, absolute difference 799. While the reviewer:author ratios are also high for India (2.6) and Korea (2.64), from the point of view of a conference organizer China stands out due to the sheer volume of submissions.

[9] We gave the reviewers a chance to request a lighter load at sign-up, and respected those quotas in our automated assignments, but there were still some over-assignments due to manual corrections of assignments by the chairs.

5.1 Reviewer training

As part of reviewer training, we prepared the following public materials (as a revision of an earlier tutorial[10], developed by Anna Rogers and Isabelle Augenstein for ARR):

- ACL'23 Peer Review Process: the general tutorial about review process for novice reviewers, that covers the basic structure of *ACL peer review process, author response, and discussion period, as well as tips for planning the time, reporting conflicts of interest and assessing whether to ask for reassignment. These materials were optional for experienced reviewers, and could be used across different *ACL venues as is.
- ACL'23 Peer Review Policies: the tutorial explaining our review form and responsible NLP checklist (§9.1), as well as our peer review policy: specific, professional reviews with scores supported by the text. Our list of reviewer heuristics such as "reject if not SOTA" currently contains 14 heuristics (continued from the original eight heuristics pioneered at EMNLP 2020 (Cohn et al., 2020)). We asked even experienced reviewers to read this tutorial. The future chairs could reuse parts of this tutorial, with necessary updates to the review form description and review policies.

Feedback. The exit survey indicates that the reviewers found the materials clear (43% respondents rated them as at 4 out of 4 and 40.5% - as 3 out of 4 on 4-point scale). One avenue of improvement suggested in many free comments was adding examples of good reviews.

We also asked the reviewers about their preferences for alternative formats, and the self-paced text-based tutorial was the majority choice (62.5% vs 13% preferring video tutorials and 9.6% preferring interactive tutorial with quizzes). But 13.4% respondents said that they would probably never be able to spend time on reviewer training, no matter what format it is offered in. This suggests that reviewer training, while valuable, will not help in all cases, and could perhaps be interpreted as an upper bound on the effect of any reviewer training.

5.2 ACL paper-reviewer matching: Area-Contribution-Language

One of the peer review issues that authors (and chairs) often complain about is "meh" reviews: the reviewer does not really find any significant problems with methodology or execution of the paper, but the overall recommendation is middling. This could be a symptom of paper-reviewer mismatch: the reviewer just is not sufficiently interested in the overall topic or approach, and hence no matter how good the paper is, it would not elicit much enthusiasm. In a recent survey (Thorn Jakobsen and Rogers, 2022) of authors, reviewers and ACs about their prior experience at NLP venues, many reviewers stated that "*the area match was right, but... the subject of the paper was not interesting to me (e.g. I would prefer another NLP task, model, or data)*" (54%), or *the paper was not asking a research question that would be interesting for me*" (45%). At the same time, over 27% of the author respondents in that survey reported that they had experience of reviews where the reviewer was not interested in the subject of the paper.

Most recent *ACL conferences and ARR work with some version of an automated paper-reviewer matching system that computes affinity scores between the abstract and title of the submission and the candidate reviewer, based on their publication history. Interestingly, the same survey by Thorn Jakobsen and Rogers (2022) found that both authors, reviewers, and ACs generally considered these scores to be the least important factor for paper-reviewer matching. Besides the limitations of the current systems, one factor here is probably the noise in the reviewer publication history data (only 75% of our reviewers indicated that their Semantic Scholar profiles were accurate enough to use for review assignments, see §4). Then there is also the inherent difficulty with establishing level of expertise on a particular topic in multi-author papers.

A traditional alternative to affinity scores, that also addresses the issue with reviewer interest, is bidding: the reviewers explicitly say which papers they would be interested in. But this process is rather laborious: for a big track, a reviewer would need to indicate their interest for hundreds of papers. It also opens up the possibility of collusion rings (Littman, 2021). In our experience, many reviewers do not even respond to bidding calls on time, which once again leads to some part of assignments being essentially random.

[10] https://aclrollingreview.org/reviewertutorial

Match by area	Match by contribution	Match by language	Review count	Review %
✓	✓	English	8996	71.36
n/a*	n/a	n/a	1052	8.35
✗	✓	English	691	5.48
✓	✗	English	558	4.43
✓	✓	✓	476	3.78
✓	✓	✗	345	2.74
✗	✓	✓	164	1.3
✗	✗	English	142	1.13
✗	✗	✓	52	0.41
✓	✗	✓	50	0.40

Table 2: The number of reviews matched to submission by different combinations of ACL (**A**rea-**C**ontribution-**L**anguage) criteria. The 'n/a' row corresponds to manual assignments by ACs, for which we do not have the match information.

Thus, we experimented with a new workflow that we dub **ACL (Area-Contribution-Language) paper-reviewer-matching**. It is a keywords-based matching process that explicitly targets three dimensions of submissions: track sub-areas (topical match), contribution types (match by focus/methodology), and target language (for submissions not focusing on English). To the extent possible, the paper-reviewer matching aimed to provide matches across all these dimensions. This approach further enabled us to provide the ACs with explanations for the specific matches (see §6.3).

Track sub-areas. Each track at ACL 2023 had an associated set of keywords describing its potential sub-areas. The goal was to describe the biggest expected sub-areas, and hopefully provide the authors with a better idea of the kind of work that the track was inviting. The full list of our keywords is publicly available in our blog post.[11] Our keywords were provided by the SACs of all tracks independently, but the future chairs may wish to take a more top-down approach to editing this list, and to ask their SACs to check that the list still describes the sub-areas for which the most submissions are expected, and the individual keywords are sufficiently clear for the authors.

Language(s). Due to the "default" status of English (Bender, 2019), submissions targeting other languages may be perceived as "niche" by reviewers. Additionally, the lack of expertise in a language may make it harder for reviewers to spot potential issues. Hence, for papers on languages other than English, we endeavoured to also maximize reviewer matches along this dimension.

Contribution types. The contribution types cross-cut tracks, and we hope they would help to decrease the amount of cases where the reviewer just fundamentally does not recognize a certain type of work (Bawden, 2019) and hence scores it down, or has unreasonable expectations (e.g. experimental results in a position paper). For example, the category of compute/data-efficiency creates a de-facto equivalent of efficiency track spread across all tracks.

Our contribution types are based on COLING 2018 classification (Bender and Derczynski, 2018), which we extended as follows: (1) NLP engineering experiment (most papers proposing methods to improve state-of-the-art), (2) approaches for low-compute settings, efficiency, (3) approaches for low-resource settings, (4) data resources, (5) data analysis (6) model analysis & interpretability, (7) reproduction studies, (8) position papers, (9) surveys, (10) theory, (11) publicly available software and pre-trained models.

Implementation. To collect the information for this kind of matching, we asked the authors at submission time to specify their preferred track (up to two), the best-matching keywords in that track (multiple selection possible, or "other" option with free text entry), the best matching contribution type(s) and target language(s). Correspondingly, at reviewer recruitment stage we asked the reviewers to fill in a form specifying their preferences for the tracks, keywords, contribution types, and the language(s) the work on which they could review. The matching itself was based on Integer Linear Programming, aiming to maximize matches across the three keyword types (with more types of matching being more valuable than

[11] https://2023.aclweb.org/blog/reviewer-assignment/

e.g. more matches only by area). As a fallback, we also retrieved Semantic Scholar profile data for the reviewers and computed the similarity between submission abstracts to the abstracts in the publication history of candidate reviewers, but this factor was given the lowest priority in the assignment strategy.

The Area-Contribution-Language matches, as well as the most similar paper of the reviewer, then also became the basis for the rationales for the match (see §6.3). The SACs were given the opportunity to selectively check and adjust the matches as described in §6.2 (although few of them did), and the ACs and SACs were able to see the rationales for the matches when considering the reviews.

From the analysis of the final 12606 reviews in START, 1052 (8.3%) did not have the match information (due to manual reviewer reassignment by the chairs, most likely emergency reviewers). Of the remaining 93.7% reviews made by our criteria, only 1.13% reviews with automated assignment were assigned based on the similarity scores from publication history, after exhausting the possible keywords-based matches in the reviewer pool. 82.9% reviews had at least one match by the type of area, 84.97% - by contribution type. Importantly for DEI efforts and development of NLP for languages other than English, we had 1167 reviews for submissions that specified at least one target language other than English – and we were able to provide a reviewer matching by (at least one) language in 63.58% such reviews.

Feedback. When asked to rate on 4-point scale how well the paper-reviewer matching worked for them, 85.5% ACL'23 reviewers rated it positively (35.7% at 4/4, 49.8% at 3/4). When asked for the kinds of mismatch, if any, 28.4% pointed at the topic, 13.7% at the methods, 10.4% at the type of contribution, 4.5% at languages, and 5.7% at other kinds of mismatch.

We conclude that Area-Contribution-Language assignments are overall a promising direction that can contribute to DEI efforts in the field and diversity of its contributions (see also §7). The matches could be further refined by (a) revising the area keywords[12], and (b) more targeted reviewer recruitment to include speakers of various languages. One of our SACs suggested providing a glossary together with the list of keywords. We also recommend investing effort into a dedicated interface for checking reviewer assignments that would enable ACs to help with reviewer assignment checks while seeing the up-to-date reviewer availability information, and highlighting the possible problems with the current assignments (such as imperfect matches, rare types of contributions or languages that may need extra attention, insufficient pool for a area or a contribution that turns out to be more popular this year).

5.3 Review issue flagging

Even with all the above efforts, we anticipated that there would still be problematic and mismatched reviews. Given that the only people with the incentive to read the reviewer guidelines and enforce them are the authors, we developed a way for them to flag reviews for specific issues, which the ACs could be given specific instructions about, and be able to address more systematically.

Unfortunately, the START system does not have an editor for the author response form or meta-review form. Hence we had to provide the authors and ACs with the list of possible issues, and ask them to specify their type and rationale in plain text form, as shown in Figure 3. As could be expected, even with a template there were many format errors. We recommend that the future conferences use a form with a multi-selector, per each reviewer.

The authors actively used this feature at ACL'23, flagging 12.9% of all reviews. This is reassuring: judging by the intensity of online discussions of peer review at each review release day, *most* reviews are bad). The frequency of various reported issues is shown in Table 3. The biggest reported problem is the heuristics such as "not novel", "not surprising", "too simple", and "not SOTA". Particularly concerning are the rude/unprofessional reviews: even though there are only 1.69%, they have the most potential to impact the mental health of the authors, and we should strive for that number to be 0.

The author-reported issues should be interpreted as a lower bound on the number of review issues, because of 100 papers were reviewed but withdrew before the final decisions. It is possible that they did because they (a) agreed with the criticism and wished to revise the paper, or (b) that they disagreed but did not see a chance to persuade the reviewers. Assuming the latter, and that all their reviews were problematic, this would raise the upper bound of problematic reviews to 15.3%. But it is unlikely that all

[12] In particular, our Language Grounding SACs indicated that their keywords should be revised and clarified.

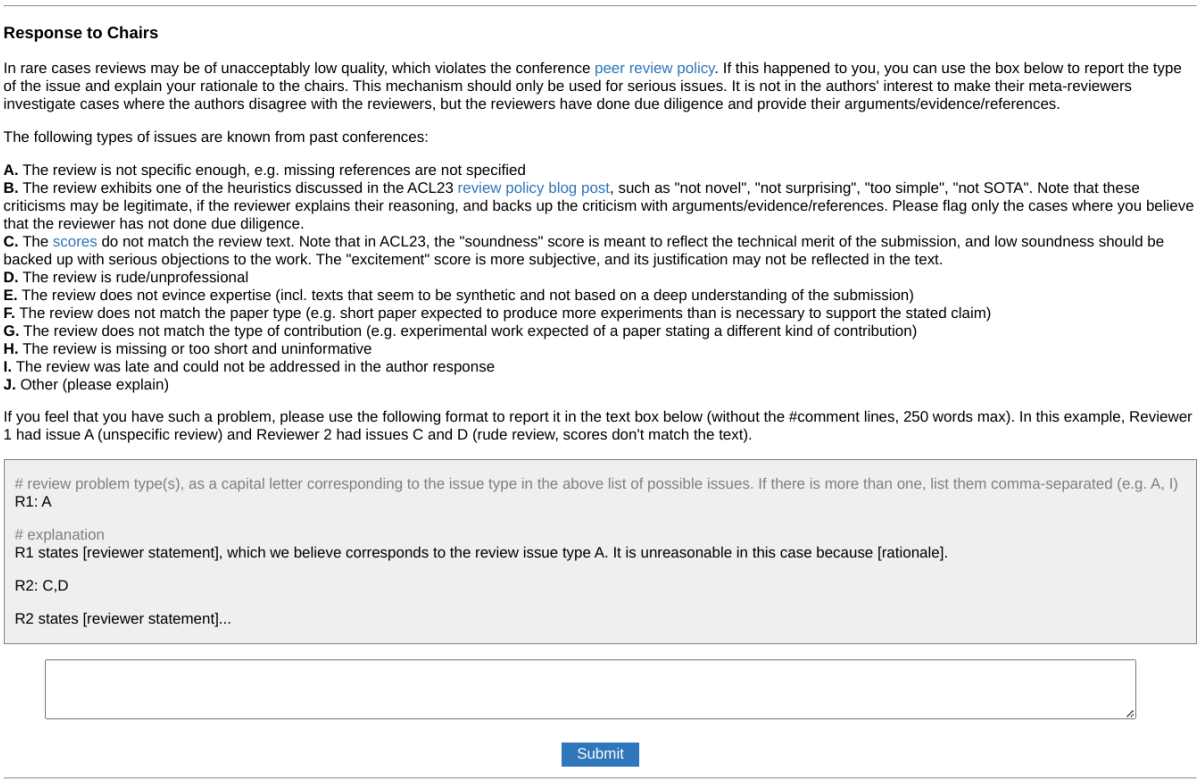


Figure 3: Review issue flagging: minimal plain-text implementation in START

withdrawn papers were of the (b) type, and the comments from ACs also suggest that many issues were not fully justified.

Feedback. When asked to rate the utility of this system at ACL'23 on 4-point scale, with 4 being the highest score, 42.1% of the authors in our exit survey rated it at 4/4, and 40.3% - at 3/4. We interpret it as overwhelming support, and recommend that this feature is maintained in the future conferences. However, the qualitative analysis of the authors' comments suggests that in some cases the ACs did not respond to the flagged issues properly, which entails the need for further training and monitoring by the SACs.

Our follow-up analysis suggests that ACs reported addressing the author-flagged issues in at least 30.59% submissions (judging by their using a similar template to Figure 3 in the "confidential notes to chairs" in the meta-review. This should be interpreted as a *lower* bound: since the interface was very clunky, it is possible that some ACs did consider the flagged issues, but did not report their actions. But, clearly, many issues were not properly addressed, and there is much room for improvement and further training of ACs. Still, given that this is the first implementation of this system, this is a promising approach and it should improve in the future.

5.4 Reviewer discussion

Similarly to most of the recent *ACL conferences, we implemented the author response period: a week during which the authors have the opportunity to read the reviews and send their response. The goal of this process is improving the quality of the reviews, and we supplemented that goal with the above new option for the authors to flag specific types of review issues (§5.3). The authors could (but didn't have to) provide a response and flag review issues; this was done for 88.3% of reviewed submissions. In 57.3% review forms the reviewers indicated that they read the response (it is possible that more did read the response but did not fill in the form).

Those comments were seen by the ACs, not the reviewers. The ACs had the *option* to initiate reviewer discussions for the cases where they saw significant disagreements, quality issues, or misunderstandings. Each paper had an associated "forum" on START, where the reviewers could communicate in an

Type of issue	*Number of reviews*	*% of reviews*
A: The review is not specific enough	272	2.16
B: Review heuristics such as "not novel", "not surprising", "too simple", "not SOTA"	678	5.38
C: The scores do not match the review text	448	3.55
D: The review is rude/unprofessional	213	1.69
E: The review does not evince expertise	542	4.3
F: The review does not match the paper type	98	0.78
G: The review does not match the type of contribution	152	1.21
H: The review is missing or too short	205	1.63
I: The review was late	12	0.1
J: Other	162	1.29

Table 3: Review issue statistics

anonymized fashion (as R1, R2, R3). The ACs were provided with instructions and suggested starter message template.

In total, out of 4559 direct submissions to ACL, 4069 had received reviews, and for 2901 out of those the ACs initiated discussions. In total, ACL review process generated 8553 messages (3879 by the ACs). However, only 2107 discussions (72.63%) had at least one response from at least one reviewer. Somewhat consistently, the discussions were overall initiated by 77.4% of all ACs. We conclude that both AC and reviewer involvement have room for improvement.

We reviewed one case of a strong paper that ended up being rejected. The AC could have been persuaded by a "champion" reviewer, and there was one such expert in the set who was surprised by the final outcome—but they did not engage in the forum discussion. We followed up with the reviewer, and they explained that since their review was already positive, they did not feel that they needed to be "on the case" anymore. We cannot establish how common this misconception is, but we would urge all reviewers to always read all reviews and author response, and when certain of the merit of a paper—to try to make sure that the AC is convinced.

6 Improving decision support for the chairs

In addition to the efforts for improving the quality of peer review (§5), we implemented the following steps for facilitating the decision support by ACs and SACs: revised SAC and AC guidelines (§6.1), guidance for assignment checking (§6.2), match rationales (§6.3), *Soundness/Excitement* scores (§6.4).

6.1 Updated SAC and AC guidelines

We updated the SAC/AC guidelines that we received from the program chairs of ACL'21 in following ways. We reformatted it to Markdown to utilize the ecosystem of GitHub (e.g., version control, asynchronous collaboration among PCs, automated deployment). The guides were built by Sphinx[13] with MyST extension[14], which enables to use Markdown and variables (making it easy to keep the consistency of dates and external URLs between SAC and AC guides and for the future chairs to adapt to their timeline). We also adjusted the existing instructions and created new instructions to incorporate everything we developed, from the new reviewer guidelines to guidelines for making recommendations. We shared the guides before the review process so that SACs and ACs can be prepared for the tasks and workloads.

Feedback. 83.3% SACS and 90.3% ACs rated the clarity of instructions at 3/4 or 4/4. Some of the free-text comments indicated a preference for shorter guidelines, but since the process is complex, and the guidelines need to serve both new and experienced chairs, there are limits to how much they can be shortened.

[13] https://www.sphinx-doc.org/
[14] https://myst-parser.readthedocs.io/

6.2 Support for checking assignments

As mentioned above, the usual workflow in large conferences is that the assignments are made automatically based on affinity scores between candidate reviewers' publication history and submissions. Usually, the automated assignments are then shown to the ACs and SACs to check manually, but this is very difficult in practice: SACs cannot process such a large volume on their own, so they need to rely on ACs. But ACs, at least on START, do not have access to the list of possible reviewers together with their current number of assignments and all their COIs, which means that even if they spot an error—it is difficult for them to identify and recommend an available alternative. Providing the up-to-date quota and COI information on all reviewers in track to the ACs is not possible in the current START platform. There are also no detailed guidelines for this step, which means that even if ACs had the reviewer information, everybody would be suggesting alternatives based on different criteria.

In our experience as SACs in previous conferences, although the automated assignments are not perfect, very few ACs actually report the problems or propose alternatives. To see whether this was widespread, we asked our SACs in the exit about whether, in their experience, the ACs asked to check the automated assignments usually recommend many changes. Only 9 of our respondents previously served as SACs in this set-up, but most of them (6/9) concurred with our experience, reporting that ACs adjust very few assignments. When asked why the ACs do not recommend more changes, 33.3% SACs stated that there are no adjustments because the ACs don't really check, 29.9%—that it happens because the automated assignments are already good enough, 29.2%—because of the difficulty with sharing up-to-date reviewer availability information with them, and 20.8%—that there are no better candidates even if the ACs check. 37.5% indicated that there are also other issues contributing to the ACs not recommending more changes.

We interpret these results as pointing to the fundamental issue of systematically sharing up-to-date reviewer availability information together with their preferences, experience, and profile information, in a way that would make it easy for the ACs to perform such checks and recommend alternatives.

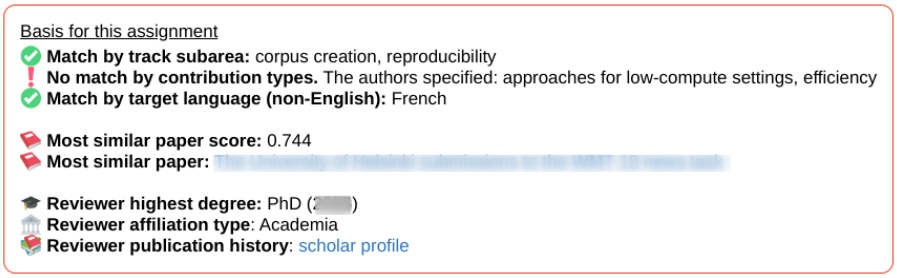
Given that the above factors make it unrealistic to adjust assignments with help of ACs, and that the volume of assignments to check was too large for SACs, we experimented with an alternative approach: since we had the "explanations" for the matches and also the quantitative information about different types of contributions, languages and area keywords, this information would make it possible for SACs to identify the types of submissions most in need of extra checks, and to focus on those. This way the workload would remain manageable, and the SACs would be able to do that while having full access to the latest reviewer availability data. To assist in this process, we developed Jupyter notebooks with quantitative analysis per track (identifying which keywords, types of contributions and languages were rare and could need extra attention)—as well as reviewer lookup functionality by preferred keywords, languages or types of contribution (or any combination thereof). This solution was better than nothing, but admittedly clunky and could be much improved.

Feedback. 66.7% of SACs stated that they believed selective checking to be overall sufficient given sufficiently strict reviewer pool criteria (although in our specific case not all reviewers in our pool were up to all SAC's standards).

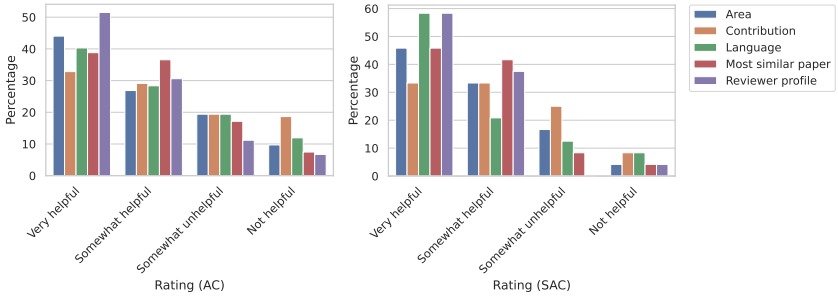
Caveat: we encountered difficulty with uploading the final automated assignments due to dynamic computation of conflicts-of-interest in START. Because of that, several hundred automated assignments had to be redone manually at the last minute. For the conferences based on START, we strongly recommend that this computation is frozen after the main part of reviewers and chairs are added to the tracks.

6.3 Paper-reviewer match rationales

Given the information for the paper-reviewer matches that we had collected (§5.2), we were able to provide the ACs with a list of rationales for each match (except for those reviewers who were added manually by the chairs, and for whom we did not have this information.) A sample "explanation" for a match is shown in Figure 4a. The idea was to provide the AC with not only the general information about the reviewer, but also what are their interests that match this submission. Importantly, we highlighted the cases where the author-stated type of contribution or language was *not* among the reviewer's stated



(a) Example of paper-reviewer match rationales. The most similar paper titles directly link to the papers (based on Semantic Scholar). For contributions and languages, the rationales either show the match, or alert to the lack of the match, so that the AC could take that into account.



(b) Chair feedback on which features of the match explanation they found the most useful.

Figure 4: Example explanation for paper-reviewer matches, and AC utility ratings for individual features displayed.

interests, which would ideally provide the AC with grounds to check potential bias against certain kinds of work.

Feedback. This feature received overwhelming support from the chairs: 87.5% SACs and 73.9% ACs rated its utility at 3 or 4 out of 4 (Figure 4b). Among the suggestions for the future improvement, the SACs suggested indicating whether the reviewer was an emergency reviewer, and how late the review was, as well as some elements of reviewer history (e.g. whether they were late for other conferences). The numerical similarity scores were less useful than the titles of the most similar papers. While predominantly the ACs were very positive about easily accessible links to reviwer profiles (Figure 4b), some ACs raised fair concerns about the effect of this feature on reviewer deanonymization: the reviewers are already visible to ACs since they need this information for chasing late reviews, but providing links to reviewer profiles increases the saliency of the reviewers' identities, and hence may by itself increase bias against, for instance, student reviewers.

6.4 *Soundness/Excitement* scores

While most of the experimental aspects of the ACL 2023 process was focused on matching reviewers to papers more effectively, a larger change visible to authors and reviewers was the introduction of two new scores on the review form to replace the *Overall Recommendation* that was previously the centerpiece of *CL review forms.

We asked reviewers for two scores: *Soundness* and *Excitement*.[15] Our goal was that any sound paper would be accepted to some ACL affiliated venue (i.e., Findings), but that the "main conference" distinction (limited by space) would be focused on the most exciting papers. Our hope was that *Soundness*, as a more specific rubric with more objective criteria, would be less noisy than a single *Overall Recommendation* score, which would help reduce the randomness of decisions. The AC guidelines had explicit instructions for how these scores should map to their recommended status.

One more factor motivating our proposal was that the *Soundness/Excitement* distinction could help with the author-reviewer communication during the author response. When a reviewer points out issues with

[15] See our definitions and rubrics for the review form and extra explanation here.

Soundness, the authors generally have a fair chance to clear any misunderstandings or issues with review quality, and the chairs are interested in this kind of discussion. The *Excitement*, however, is subjective, and the authors do not have a fair chance to convince reviewers that their general views or research agenda are wrong. The *Soundness/Excitement* distinction helps to focus the response on the *Soundness* issues, and hence have a more productive discussion.

Feedback. Judging by the exit surveys, this change was overall well received: over 80% of the chairs, reviewers and authors either expressed support or did not object to this change. 38.1% authors, 35.1% reviewers and 29.9% ACs indicated that while the idea was good, it could be better executed. Among the named issues was the clarity of communication about what these scores meant, the difference in granularity (our scale for *Excitement* had 9 points, and *Soundness* only 5), and the wording could be adjusted to remove the semblance to *Overall recommendation* score. We made these recommendations to the program chairs of EMNLP 2023, who decided to keep this system.

From the communication with the authors who expressed dislike for this system, our impression is that one of the factors here is the mistaken impression that the final decisions are overall based on scores, and the papers with similar scores should be guaranteed the same outcome—whereas in reality the chairs know that scores can be noisy and miscalibrated, and hence the final decisions are made on case-by-case basis, with the full view of the reviews and meta-review, and also taking into account the acceptance quotas and their editorial priorities.[16] The *Soundness/Excitement* scores were rather intended to make it harder for the chairs to just sort by the scores.

7 What Factors Contribute to ACL Peer Review Outcome?

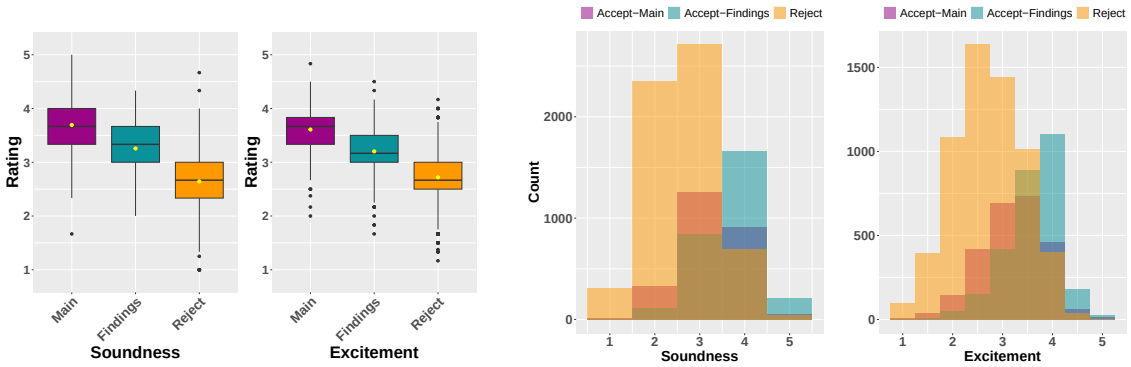
Here we present the results of statistical analysis of ACL'23 data, with the goal of explicating what factors contributed to the final decisions and to the quality of individual reviews. We hope that this process both improves the transparency around chair decision-making, and highlights the potential biases and points of improvement for future conferences.

For the new authors, we should explain the general process for the acceptance decisions at ACL'23. First, the reviewers contribute their reviews. At the author response the authors see the reviews and have an opportunity to respond: a process mostly intended to clarify any misunderstandings (we disallowed submitting new results). Then the ACs initiate the reviewer discussion, with the goal to clarify misunderstandings and improve the quality of the reviews. Based on the final reviews and their own expertise, they write the meta-reviews and make recommendations for acceptance (Main track or Findings) or rejection. They are *not* concerned with the acceptance quotas. Their recommendations and meta-reviews (as well as reviews and author response if necessary) are then considered by the SACs, who have the constraint of the target acceptance quota (which we set at about 22% for the main track and 35% for Findings). Their decisions are based on three main factors: meta-reviews, quotas, and editorial priorities (with case-by-case consideration as needed). If they run out of their quota, they may additionally rank more papers by priority that may be accepted to main/track Findings if there is space (e.g., because some tracks did not use their quota fully). The final step is that the program chairs confirm the SAC decisions, and try to fit in as many papers of the ranked "maybes" as possible. In our case, that resulted in accepting more Findings papers than we originally planned based on prior conferences.

7.1 Review Scores: Overall Distribution

We start by exploring the overall distribution of the new *Excitement* and *Soundness* scores (described in §6.4) and how they mapped to the three possible decision outcomes (Rejection, acceptance to the Main track, or Findings). Both *Excitement* and *Soundness* are ordinal variables, and we use the mean as a rough estimate of the central tendency. Figure 5a shows that for both scores the means are higher for main track than for Findings, and for Findings they are higher than for rejections. For *Excitement* this is fully in line with our instructions to the chairs. For the main track, this suggests that higher (above 3) *Soundness* scores

[16] This is a general problem, and we imagine this would have also happened in the case of an *Overall recommendation* score. The drawback of the *Soundness* plus *Excitement* system is that less noisy decision cutoffs make outliers more salient



(a) Distribution of mean scores

(b) Distribution of individual scores

Figure 5: *Soundness* and *Excitement* scores per acceptance status

	Findings Coeff	Main Coeff	Findings SE	Main SE
(Intercept)	-1.48	3.77	0.79	1.43
Soundness Mean	0.71	0.76	0.22	0.37
Excitement Mean	0.61	0.03	0.23	0.42
AC Recommendation (L)	2.66	4.50	0.50	0.94
AC Recommendation (Q)	-1.16	-0.05	0.43	0.81
AC Recommendation (C)	-0.04	0.10	0.31	0.58
AC Recommendation (^4)	0.04	-0.27	0.19	0.37
SAC Recommendation (L)	5.84	28.26	0.47	0.71
SAC Recommendation (Q)	-1.06	13.59	0.34	0.77
SAC Recommendation (C)	1.18	7.82	0.60	0.82
SAC Recommendation (^4)	1.52	4.48	0.64	0.74

Table 4: Coefficients and Standard Errors (SE) for the Multinomial Logistic Regression Model predicting the final acceptance decisions given the mean scores and AC/SAC recommendations. Each row corresponds to a predictor in the model, with separate coefficients reported for each level of the outcome variable (Findings and Main). The 'L', 'Q', 'C', and '$\hat{4}$' subscripts for AC_ordinal and SAC_ordinal represent linear, quadratic, cubic, and quartic polynomial terms, respectively, reflecting the assumed shape of the relationship between these ordinal predictors and the log-odds of the outcomes.

also played a role in main vs Findings decisions, although the difference is less than between Findings and rejection. The overall score distribution is shown in Figure 5b.

7.2 Factors Impacting the Final Acceptance Decisions

7.2.1 Reviewer Scores and Chair Recommendations

To establish the odds of a paper being accepted into Findings or the Main track vs it being Rejected, based only on reviewer and chair recommendations, we fit a multinomial log-linear model with `multinom()` function from the NNET package in R (Venables and Ripley, 2002).[18] The dependent variable (DV) is the *Outcome* coded as a three-layer categorical variable (Main track, Findings, or Reject) with Reject being set as the reference level. The independent variables (IVs) are *AC Recommendation* (ordinal), *SAC Recommendation* (ordinal), mean *Soundness* score (interval), and mean *Excitement* score (interval).[19] The analysis is performed on the papers submitted directly to the conference as the ARR submissions were reviewed through a different process and had different scores. The model coefficients are shown in Table 4. The model is a good fit for the data with McFadden's pseudo-R^2 of 0.777 (McFadden, 1973).[20]

[17] Signif. codes: '$p < 0.001$' '***', '$p < 0.01$' '**', '$p < 0.05$' '*', '$p < 0.1$' '.', '$p > 0.1$' ' '.

[18] While ordinal regression would be more fit to represent the ordinal order of the possible outcome (Main track > Findings > Reject) we use the multinomial model as it does not have the proportional odds assumption.

[19] Note both, the *Excitement* and *Soundness* are ordinal variables. Here, we employ the mean to obtain a rough estimate of the central tendency.

[20] Please note the pseudo-R^2 for logistic models cannot be directly interpreted as the proportion of variance explained as in linear models. Nevertheless, the high value observed here signifies a good fit to the data. We also report Cox and Snell

	LR Chisq	Df	Pr(>Chisq)	
Soundness Mean	10.88	2	0.0043	**
Excitement Mean	9.67	2	0.0080	**
AC Recommendation	209.71	8	0.0000	***
SAC Recommendation	1438.12	8	0.0000	***

Table 5: Type III Analysis of Deviance for Multinomial Logistic Regression in Table 4.[17]

To obtain the significance values for each IV (Table 5), we use the ANOVA() function in R on the fitted model (Type III Anova). As expected, all four IVs are significant (p<0.05) but at different levels. The *SAC Recommendation* ($\chi^2(8) = 1438.12$, $p< 0.001$)[21] and *AC Recommendation* ($\chi^2(8) = 209.71$, $p< 0.001$) significantly predict the *Outcome* with the *SAC Recommendation* appearing to be a better predictor (as expected, since *AC recommendation* are made without regards to the acceptance quotas). The mean *Soundness* score ($\chi^2(2) = 10.88$, $p= 0.0043$) and mean *Excitement* score ($\chi^2(2) = 9.67$, $p = 0.0080$) are also significant at p<.05.

To establish the exact contributions of mean *Soundness* and *Excitement* scores to acceptance decisions for the Main track and Findings, we can look at Table 4 again. Note that since it is a multinomial regression model, the coefficients indicate an increase in log odds rather than directly interpretable odds (for which the coefficients need to be exponentiated). The "Findings Coeff" and "Main Coeff" correspond to the log-odds of being accepted into the Findings and Main track as opposed to being rejected.

Soundness. In the case of the mean *Soundness* score the coefficient is positive for both Findings (0.71) and the Main track (0.76). This means that for one unit increase in the mean *Soundness* score the log-odds of being accepted as opposed to being rejected increase by 0.71 for Findings and 0.76 for the Main track. By taking the exponential of these values, we see that for one unit increase in the mean *Soundness* score the odds to be accepted increase 2.03 times for Findings and 2.14 times for the Main track.

Excitement. Similarly, both coefficients are positive for the mean *Excitement* score for both Findings (0.61) and the Main track (0.03). This means that for one unit increase in the mean *Excitement* score the log-odds of being accepted vs rejected increase by 0.61 for Findings and 0.03 for the Main track. By taking the exponential of these values we see that for one unit increase in the mean *Excitement* score the odds of being accepted increase 1.84 times for Findings and 1.03 times for the Main track. While the values are still positive, this increase is much lower[22] than for the mean *Soundness* scores, especially for the Main track. The overall distribution of these scores per acceptance status is shown in Figure 5b.

AC Recommendations. Since *AC Recommendation* is an ordinal variable, it is coded using polynomial contrast, so the L indicates linear effect, Q a quadratic effect, C a cubic effect, and so on. Here we look mostly at the linear effect since it has a direct (linear) effect on the outcome. We see that both coefficients are positive, indicating that with an increase of one unit, the log-odds of being accepted vs being rejected increase by 2.66 units for Findings and 4.50 units for the Main track. By taking the exponential of these values we see that one unit increase in *AC Recommendation* corresponds to a 14.30-fold increase in the odds of being accepted into Findings (vs being rejected) and 90.02-fold increase in the odds of being accepted into the Main track (vs being rejected).

SAC Recommendations. *SAC Recommendation* is also an ordinal variable, hence we see the same types of coefficients. However, the magnitude of the SAC's decision appears to be much greater with a greater effect on the final outcome. With one unit increase in *SAC Recommendation* the log-odds of being accepted vs being rejected increase by 5.84 units for Findings, and 28.26 units for the Main track.

pseudo-R^2=0.794 (Cox and Snell, 1989) and Nagelkerke pseudo-R^2=0.913 (Nagelkerke, 1991).

[21] χ^2 denotes likelihood ratio chi-square statistic.

[22] This latter finding seems counter-intuitive, given that our AC guidelines stressed that Findings is a venue for all sound work, while "sound& exciting" would be the basis for recommendations to the main track—but even among the papers accepted to the main track 39% have at least one "negative" *Excitement* score (Figure 7b). At the same time, even among the Findings papers, only 49% have predominantly negative *Excitement* ratings, so there is a preference for at least *some Excitement*. This could be related to the confusion about the meaning of the scores in the initial iteration (see subsection 6.4).

	LR Chisq	Df	Pr(>Chisq)	
Paper Type	12.47	2	0.0020	**
Review Issues	43.61	2	0.0000	***
Preprinted	47.96	2	0.0000	***
Previous Submissions	4.38	2	0.1120	
Languages Number	0.57	2	0.7528	
Languages not only English	3.53	2	0.1711	
Contribution: Efficiency	1.18	2	0.5540	
Contribution: Resource	4.34	2	0.1139	
Contribution: Reproduction	16.59	2	0.0002	***
Contribution: Theory	7.70	2	0.0213	*
Contribution: Software	19.62	2	0.0001	***

Table 6: Type III Analysis of Deviance for Multinomial Logistic Regression, predicting submission *Outcome* (Main, Findings, Reject) conditioned on the variables listed in the table.[24]

Converting these values to their exponentials, we see that one unit increase in *SAC Recommendation* corresponds to a 343.78-fold increase in the odds of being accepted into the Findings (vs being rejected) and a massive increase of 1.88×10^{12} for the odds of acceptance into the Main track (vs being rejected).

The model hence shows that the SAC recommendation is a much stronger predictor than the AC recommendation, which helps to explain why it is possible for a paper to be rejected even with a positive meta-review. AC recommendations are made without regards to the acceptance quotas, and SACs necessarily have to override them in many cases.

7.3 The Impact of Other Submission Properties

There are many properties of submissions that could systematically make a difference to their final outcome. In this section we investigate the possible effect of the type of contribution, the target languages, whether the reviews were problematic (as reported by the authors), and whether the paper was available as a preprint. To establish the importance of these factors, we fit another `multinom()` model, similarly to what we did in Table 4, and obtain the significance levels for each variable using Type III Anova. While the ordinal model would potentially better preserve the natural order of the final outcome (rejection being the worst and acceptance to the main track being the best outcome), the fitted model violated the assumptions of the ordinal model.

Since this model does not include strong predictors such as reviewer scores and chair recommendations, the fit of this model is relatively poor[23] compared to the model in Table 4, which has a McFadden's pseudo-R^2 of approximately 0.80 (indicating a substantial improvement over the null model). In contrast, this model has a McFadden's pseudo-R^2 of approximately 0.01, suggesting that it barely improves upon the null model. Nevertheless, this model can still be used to establish the individual contributions of the submission-level properties, which likely interact in complex ways in the scores and recommendations. Statistically significant factors are also not necessarily strong predictors by themselves.

The results of this experiment are shown in Table 6. According to this analysis, the following factors have a statistically significant impact on submission outcome: low-quality reviews, preprinting, short/long paper type, and three types of contributions (software, reproduction, and theory).

To also assess the relative importance of our predictors in forecasting the final outcome, we employed a Random Forest algorithm (Liaw and Wiener, 2002). The results are shown in Figure 6. The most crucial predictor was *Review Issues* (i.e., author complaints about reviews[25]) with a Mean Decrease Gini value of 46.09. This suggests that this predictor played the most significant role in reducing the Gini impurity, and therefore, in improving the precision of our model. The second factor with the biggest Mean Decrease Gini is *Preprinting* (22.84). This analysis does not state the absolute importance of any factor (e.g., that

[23] Its 3-class accuracy is 52%, vs 90% for the model shown in Table 4. This is the accuracy of the model on the withheld test set when the model is fitted with 70% of the data. The accuracy of the model on all data is about 1% higher.

[24] Signif. codes: '$p < 0.001$' '***', '$p < 0.01$' '**', '$p < 0.05$' '*', '$p < 0.1$' '.', '$p > 0.1$' ' '.

[25] The number of author complaints likely reflects (at least) two factors: the reviews that were truly problematic, and simply negative reviews since the authors are more likely to complain about those. In the latter case the leading cause for rejection is the negative review.

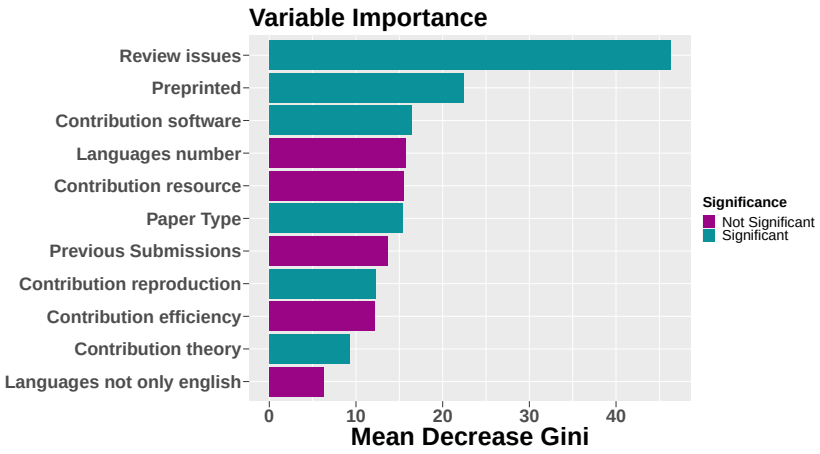


Figure 6: The importance of predictors in predicting the *Outcome*, ranked by mean decrease in Gini impurity. Predictor significance is indicated by color, with dark purple for not significant and dark green for significant predictors as per levels of significance indicated in Table 6.

Contribution type	*% submissions*	*Match*	*Mismatch*	*Match-Mismatch*
Efficiency	9.62	50.27	46.56	3.71
NLP engineering experiment	61.5	46.66	47.33	-0.67
Software and pre-trained models	12.14	56.75	45.56	**11.19**
Data resources	19	49.25	46.37	2.88
Data analysis	10.48	48.14	46.78	1.36
Reproduction studies	2.08	66.25	46.51	**19.74**
Approaches for low-resource settings	18.22	49.79	46.28	3.51
Surveys	1.64	44.44	46.96	-2.52
Interpretability	25.29	51.8	45.27	6.52
Theory	3.8	56.85	46.53	**10.32**
Position papers	2.57	53.54	46.74	6.8

Table 7: Acceptance rate among direct submissions that were reviewed and considered for acceptance, with (*Match*) and without (*Mismatch*) given contribution types. The average acceptance rate in this pool is 46.92%.

Preprinting increases the chances of acceptance by X%), and we are not claiming that these effects are independently large—but they do appear to be statistically significant. We will discuss these factors further: short/long papers in §7.3.1, contribution types in §7.3.2, review issues in §7.5.5, preprints in §7.5.7.

7.3.1 Short/long papers

Short papers have had significantly lower acceptance rates at most recent *ACL conferences. To mitigate that, we highlighted the problem in the reviewer instructions, had a separate *Soundness* formulation for short papers, and asked the SACs to consider the short and long papers separately, with their own target acceptance quotas. Despite all that, the significant effect of paper type (Table 6) is obvious: the long papers had 23.50% acceptance rate to main track vs 16.53% for short, and for Findings, the rate was respectively 41.89% vs 35.58%. The core reason seems to be that the source reviewer scores are systematically lower, despite all calls to not expect 120% thoroughness of short papers.

7.3.2 Types of contribution

We were pleasantly surprised to find a significant positive effect for the contributions of theory, reproductions, and pre-trained models and software (Table 6). The two latter types are in line with the findings by (Magnusson et al., 2023) who report that reproducibility efforts are rewarded. This effect is also visible from simply considering the differences in acceptance rates for papers with and without these contribution types, shown in Table 7. In fact, the "average" acceptance rate of 46.92% is the closest to the most "mainstream" type of contribution (NLP engineering experiment, 61.5% submissions) – and all other contribution types except surveys have the acceptance rate at least slightly higher than that.

Submissions subset	*Contribution type*	*% submissions*	*Match*	*Mismatch*	*Match-Mismatch*
Resources & Evaluation	Resource	5.48	48.39	48.21	0.18
All tracks without Resources & Evaluation	Resource	94.52	49.48	46.34	3.14
Interpretability and Analysis of Models	Interpretability	4.89	52.69	57.14	-4.45
All tracks without Interpretability	Interpretability	95.11	51.61	45.18	6.43

Table 8: Acceptance rate among direct submissions inside and outside tracks that targeted a resources and interpretability contributions, with (*Match*) and without (*Mismatch*) given contribution types. The average acceptance rate in this pool is 46.92%.

		Accepted papers only		Rejected papers only		All papers	
		%	$\alpha_{[CI]}$	%	$\alpha_{[CI]}$	%	$\alpha_{[CI]}$
Ordinal	*Soundness*	20.72	$0.093_{[0.047,0.137]}$	17.68	$0.116_{[0.076,0.156]}$	19.10	$0.318_{[0.294,0.340]}$
	Excitement	12.68	$0.120_{[0.075,0.169]}$	10.65	$0.134_{[0.094,0.173]}$	23.23	$0.311_{[0.287,0.334]}$
Categorical	*Soundness*	77.28	$0.032_{[-0.052,0.112]}$	37.39	$0.092_{[0.064,0.119]}$	53.80	$0.221_{[0.194,0.248]}$
	Excitement	37.11	$0.087_{[0.055,0.120]}$	49.60	$0.074_{[0.039,0.114]}$	43.74	$0.233_{[0.212,0.255]}$

Table 9: Inter-reviewer agreement on soundness and excitement scores, measured as raw % agreement (%) and Krippendorff's alpha (α) with 95% confidence interval [CI].[26] We consider only direct submissions to ACL'23 that were fully reviewed, and for which the final decisions were made: 3847 in total, 1805 "accept" (to either Main track of Findings), and 2042 "reject".

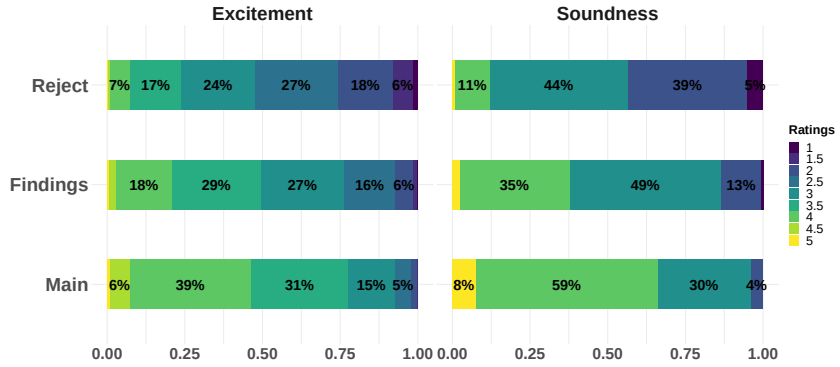
The *lack* of a visible disadvantage in acceptance rates for non-mainstream types of contributions is a very positive finding. Consider the case of efficiency-oriented papers: they did not have a dedicated track, but their acceptance rate was not lower (and even a bit higher) than for the average in the pool (where the majority of engineering-oriented submissions focuses on performance). In effect, *every* track was an efficiency track, allowing both access to the area expertise and reviewers with interest in this type of contribution. We cannot establish to what extent this is due to Area-Contribution-Language matching or an overall increased interest in the need for efficient NLP solutions. But as long as such contributions are in the minority, we would recommend ensuring the matches by this criterion.

A complication for our analysis arises for two contribution types that also had large associated tracks: resources and interpretability. In this case, it is possible that the lack of difference in acceptance rate is due to the extra effort of ensuring the reviewers with matching interests through the track mechanism. To check for that, we compare the acceptance rates for these types of contributions inside and outside of the dedicated tracks (Table 8). We find that in all cases the match between tracks and contribution types yields a 3-6% increase above the average acceptance rate of 46.92%. An interesting case is interpretability and model analysis, which has a 4.45% higher acceptance rate *outside* of its dedicated track (probably indicating an appreciation for papers that perform analysis in addition to some other type of contribution).

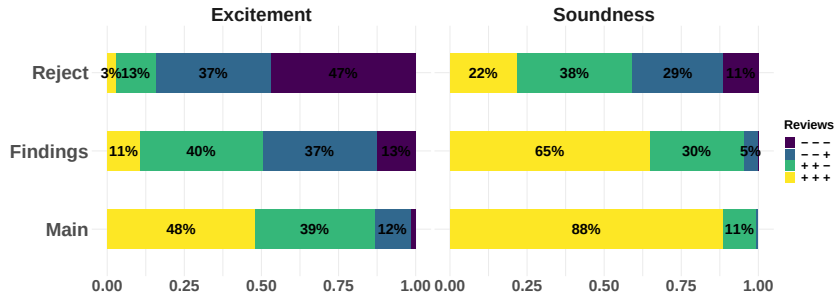
7.4 How Much do ACL Reviewers Agree?

The issues with consistency of peer review were recently highlighted in the ML community by the two NeurIPS experiments (Price, 2014; Cortes and Lawrence, 2021; Beygelzimer et al., 2021). By treating peer review as an annotation problem (Rogers and Augenstein, 2020), we can apply the existing methodology for analyzing inter-annotator agreement (IAA). We consider three reviewers (annotators) per paper, discarding the rare cases of 4 reviews (from emergency assignments). We compute Krippendoff's α (Krippendorff, 2011) on the *Soundness* and *Excitement* scores (Table 9). We treat these scores as ordinal data. We also experiment with mapping both scores to binary "positive/negative" categories (3–5 > "sound" for *Soundness* and 3.5–5 > "exciting" for *Excitement*, since the borderline scores were 2 for *Soundness* was 2 and 3 for *Excitement*).

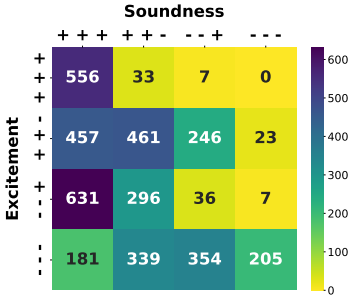
[26] "Ordinal" refers to the α coefficient computed using raw scores treated as ordinal variables. The percentage agreement for *Soundness* was computed using the raw scores (5-point scale). In order to match the scale length the percentage agreement for *Excitement* was computed on the rounded scores (i.e., 3.5 was treated as 4.0, etc.). "Categorical" denotes scores converted into either positive or negative decisions based on the given threshold (3.0 for *Soundness* and 3.5 for *Excitement*).



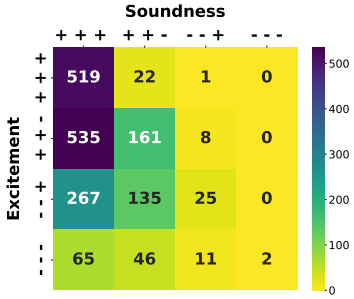
(a) Ratio of review scores per acceptance status



(b) Ratio of positive (+) and negative (-) score combinations per acceptance status



(c) Total number of submissions with different combinations of positive (+) and negative (-) scores



(d) Number of accepted submissions with different combinations of positive (+) and negative (-) scores

Figure 7: Review scores vs acceptance outcome. "Positive" scores (+) refer to the above-borderline scores (*Soundness* >=3, *Excitement* >=3.5), and "negative" (-) - to the number of scores below borderline.

Consistent with the general perception of inconsistency in peer review, α shows a level of IAA that seems far too low (the rule of thumb is that "substantial" agreement is in the range of 0.6-0.8 (Artstein and Poesio, 2008; Paun et al., 2022)). However, **the raw agreement for the accepted papers (in the categorical view, i.e. as sound/unsound, exciting/unexciting) is almost twice higher for *Soundness* than for *Excitement***. We interpret this as an indication that although the scores are still noisy, it helps to ask more specific questions with more objective criteria. The much lower raw agreement on the *Excitement* is also in line with our point that this is overall a less relevant direction for the author response and reviewer discussion. Arguably we do not even want a high agreement on *Excitement*: everybody interested in the same thing could indicate that the field is ossifying and stagnating.

As a sanity check, we also analyzed IAA for the raw reviewer scores of EMNLP 2022 and EACL 2023. Both of these conferences used a single "overall recommendation" score, formulated differently for short and long papers. In EMNLP 2022, for 3092 observations for 3 reviewers (discarding R4 data), with scores treated as ordinal data, we got α 0.316 for the short papers, 0.31 for long, and 0.318 for the whole distribution – which is almost exactly the same as our α for both our scores (in the ordinal case). In EACL 2023, for 1121 subjects for 3 reviewers we got α 0.317 for the short papers, 0.34 for long, and 0.348 for the whole distribution.

A related question is "what kind of disagreements do we actually have?" Figure 7a shows the distribution of individual score values for all papers in a given acceptance status, which suggests that even papers accepted to the main conference had some very negative reviews. Figure 7b breaks down the scores into "positive" (*Soundness* >= 3, *Excitement* >= 3.5) and "negative", and considers the combinations of three reviews as "all positive" (+ + +), "all negative" (- - -), "2 positive, 1 negative" (+ + -) and "2 negative, 1 positive" (- - +). We can see that despite disagreements on the exact scores, the papers accepted to the main track have a high ratio of "positive" review combinations for *Soundness* (88%, only 11% papers with one negative *Soundness* score). But for *Excitement* our SACs accepted to the main track 39% papers with one negative *Excitement* score, and 37% papers with a single "champion" reviewer. For Findings, they even accepted 37% papers which only 1 reviewer was excited about. Figure 7c shows the total number of submissions with various combinations of positive and negative *Soundness* and *Excitement* scores, and Figure 7d shows the same categories, but with the number of accepted papers with that score combination.

Our data indicates that despite noisy scores and high disagreement, the mechanism of ACs and SACs does "rescue" many papers with one negative review, and at least the raw agreement does improve for the more specific *Soundness* score. Judging by the community feedback (§5), in this first implementation there was a lot of confusion about what the scores meant, and we expect that in future iterations the agreement could improve further.

7.5 Analysing Reviews and Review Scores

In this section, we take a step back from the final acceptance decisions and look only at the individual reviews and their scores, rather than the final outcome of the submission.

7.5.1 Do the Area-Contribution-Language matches impact reviewer scores?

To answer this question, Figure 8 shows the distributions of the individual reviewer scores for *Soundness*, *Excitement*, reviewer *Confidence*, and *Reproducibility* for all cases where the reviews were or weren't matched by the area, contribution type, or language. The biggest visible impact is in reviewer *Confidence*, where the contributions are not matched by area: the ratio of reviews with high scores (4+) is decreased by about 14%. A worrying observation is that there is a 5% *increase* in high *Confidence* scores for the submissions where the reviewer is *not* matched by language and could be expected to feel less rather than more confident. We also observe an 11% increase in *Soundness* ratings 3+ from reviewers matched by language vs those mismatched, and 7% in *Reproducibility*.

7.5.2 Do the Area-Contribution-Language matches impact the reviewer activity?

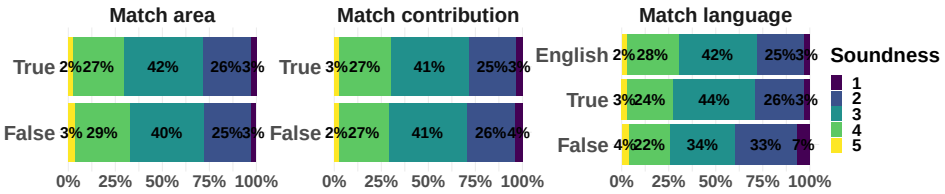
To establish whether Area-Contribution-Language matching had any effect on reviewer activity, we counted the reviewers as "active" if they had at least one forum message or more than one review edit. The distributions of active/inactive reviewers that are/aren't well-matched to submissions by Area-Contribution-Language criteria are shown in Figure 9. At a glance, there are a lot more matched & active reviewers, but since generally a lot more reviewers were matched than mismatched (see Table 2), we would generally expect that to be the case even by chance.

To establish whether there are any statistically significant effects, we first fit a generalized linear model (GLM) using the `glm()` function in R.[27] The dependent variable was binary (the activity of the reviewer). The predictors were a contribution match (binary variable), a studied language match (three-layer categorical variable),[28] and an area match (binary variable), all of which were treated as categorical variables (at least one matching keyword of the correct type). The link function was logit, corresponding to a binomial distribution of the response variable (logistic regression).

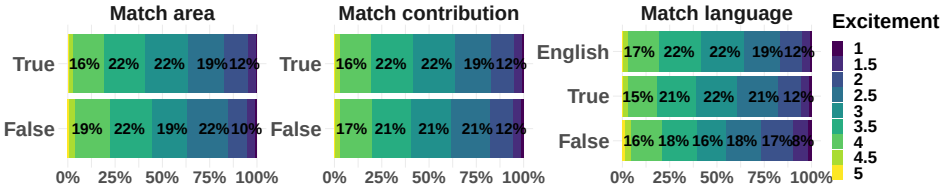
[27] To validate the assumptions of the GLM, we examined the variance inflation factors (VIFs) using the `vif()` function in R to assess multicollinearity among predictors. The VIFs were all close to 1, suggesting that multicollinearity was not a concern. We also visually inspected residual plots to assess the model fit and did not find any obvious deviations from homoscedasticity or linearity.

[28] For the language we consider three categories: (1) non-English language match, (2) non-English language mismatch, and (3) match only by English; under the assumption that all reviewers will be familiar with English.

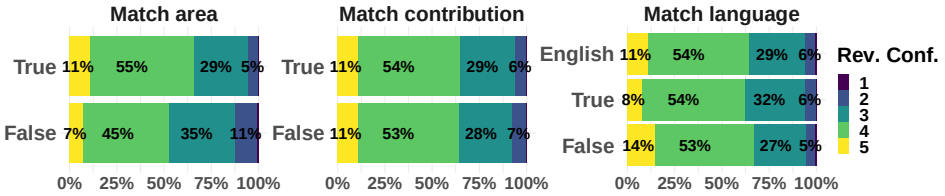
[29] Signif. codes: '$p < 0.001$' '***', '$p < 0.01$' '**', '$p < 0.05$' '*', '$p < 0.1$' '.', '$p > 0.1$' ' '.



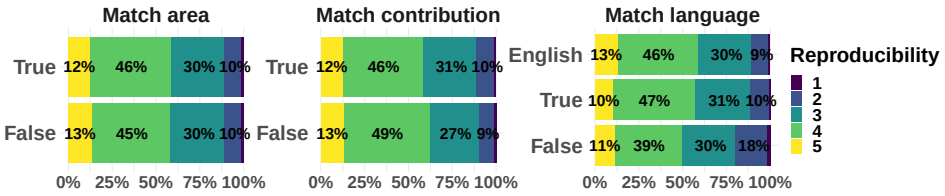
(a) Soundness



(b) Excitement



(c) Reviewer confidence



(d) Reproducibility

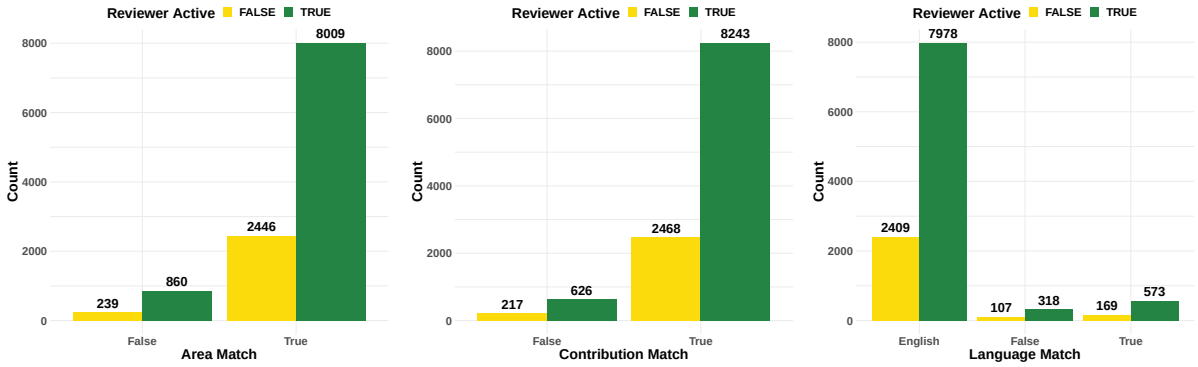
Figure 8: Area-Contribution-Language Matches impact on reviewer scores. In each plot, True/False refers to the reviews where the submissions were/weren't matched by area, contribution or language.

The results of the GLM (see Table 10) suggest that contribution match is a significant predictor of the reviewer's activity ($\beta = 0.16$, SE = 0.08, $z = 1.97$, $p = 0.048$). Since the estimates relate to log-odds we consider the exponential of the reported value (1.178) which suggests that the odds of the reviewer being active when the contribution type is well-matched are 1.178 times higher than when the contribution does not match the reviewer's expertise. The remaining variables, that is language match and area match, are not significant predictors in this model ($p > 0.05$).[30]

Finally, we considered the language match as a binary variable, excluding English language papers. We conduct a Chi-square test (χ^2) to examine the association between the language match (excluding English) and reviewer activity Table 11. The test reveals no significant association between the language match and reviewer activity ($\chi^2(1)$=0.73432, $p = 0.3915$). The chi-square test was performed using Pearson's Chi-squared test with Yates' continuity correction with the `chisq.test()` function in R.

We conclude that of the Area-Contribution-Language matching rubrics, only the contribution type contributes to improvement in reviewer activity. Although the effect is modest (1.178 times increase in likelihood of reviewer activity), given that reviewer activity post-submission is very important, and its level needs to be improved (§5.4), we would urge the future chairs to consider this criterion in the assignments. It also provides a quick and interpretable way to consider the variety of the types of work

[30] McFadden's pseudo-R^2 of the model is 0.0008231973, which is very low. This suggests that our model does not explain much of the variability in the data. However, it is important to note that in the context of generalized linear models, the interpretation of pseudo-R^2 is not as straightforward as it is in ordinary least squares regression. The pseudo-R^2 is not necessarily a measure of the proportion of variance explained by the model in the data. Instead, it is a measure of the likelihood improvement per observation relative to the null model. Despite the low pseudo-R^2, our model could still provide valuable insights into the relationships between the independent variables (match type) and the reviewer's activity.



(a) Match by area (b) Match by contribution (c) Match by language

Figure 9: Area-Contribution-Language matches vs reviewer activity. In each plot, True/False refers to reviews where the reviewers weren't matched to the submission by area, contribution or language

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.1511	0.1012	11.38	0.0000	***
Match Contribution (True)	0.1638	0.0830	1.97	0.0484	*
Match Language (False)	-0.1076	0.1142	-0.94	0.3461	
Match Language (True)	0.0114	0.0921	0.12	0.9015	
Match Area (True)	-0.1151	0.0786	-1.46	0.1432	

Table 10: Generalized linear model (GLM) estimates for predicting reviewer activity using match categories. Each row represents a different predictor.[29]

Test	Chisq	df	p-value
Pearson's Chi-squared (Yates' correction)	0.73432	1	0.3915

Table 11: Results of Pearson's Chi-squared test with Yates' continuity correction for the effect of language match (excluding English) on the reviewer's activity

that are being submitted, and to provide extra attention to the assignments for the non-mainstream kinds of work.

7.5.3 Do reviewer confidence scores reflect their experience?

START profiles contain self-reported reviewer experience labels ("never", "first time", "3 or fewer events", "4 events and more". We explored the relationship between this data and reviewer *Confidence* scores but found no strong effect. We do observe a small (about 4%) increase in the volume of 4+ *Confidence* scores for the most experienced reviewers, and it's significant according to the ordinal logistic regression model[31]. But the effect is quite small, and judging by this data we don't recommend relying on confidence as a proxy for reviewer experience. Moreover, we observe no relation between this reviewer experience data and the number of review issues reported by the authors. This is a rather depressing finding from the perspective of reviewer training, and we hope that it is rather due to START profiles not being updated by the reviewers.

7.5.4 Do the reviewer scores correlate with length of the reviews?

The ACL review form had the following text input fields: summary, reasons to accept, reasons to reject, questions to the authors, missing references, suggestions&typos, and confidential notes to the chairs. We roughly estimated the length of these inputs by splitting on the whitespace, and computed Spearmans correlation (Spearman, 1987) between these variables and reviewer scores for *Soundness*, *Excitement*,

[31] We fit model in R using the `polr()` function from the `MASS` package (Venables and Ripley, 2002) with reviewer's confidence as an ordinal DV and experience as a three-layer categorical IV. We compare this model to an intercept-only model using the `Anova()` function. While the difference between these models is significant, McFadden's *pseudo-*R^2 is extremely low (4.247533×10^{-4}).

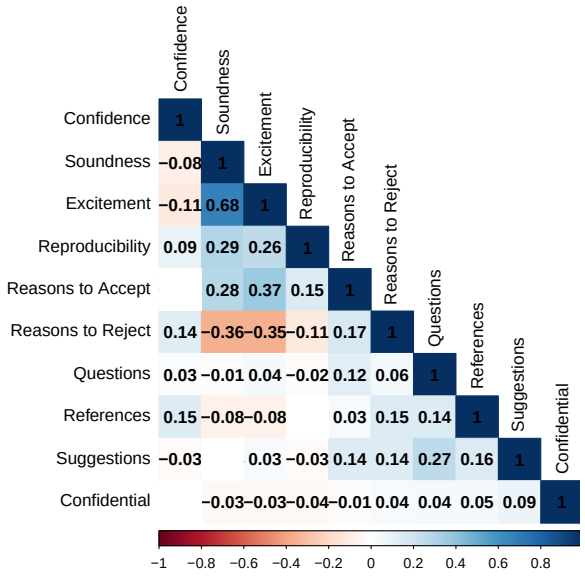


Figure 10: Spearmans' correlation between reviewer scores, confidence, and the length of review text fields. The insignificant correlation was left blank (p>0.05).

Confidence, and *Reproducibility*. The results are shown in Figure 10.

As could be expected, we observe a significant negative correlation (-0.35-0.36) between the length of *Reasons to Reject* and both *Soundness* and *Excitement* scores, and the opposite trend for the *Reasons to Accept* (0.28-0.37). Interestingly, the length of *Reasons to Accept* also correlates positively with the *Reproducibility* score, indicating that the community appreciates this factor (0.15). *Confidence* has a similar correlation with the length of missing references. Finally, there is a high correlation between the length of "questions to the authors" and "suggestions", indicating that the reviewers who engage with the submission deeply use both of these fields.

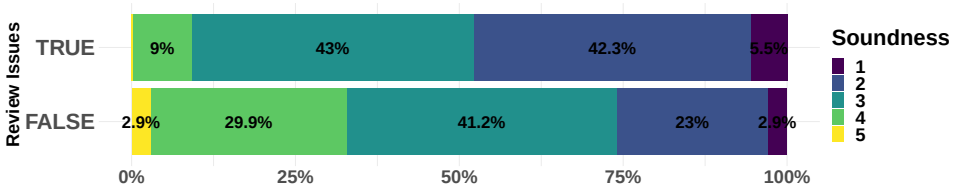
The highest positive correlation is between our *Soundness* and *Excitement* scores[32] (0.68), which is in line with the intuition that unsound work would probably not be found exciting either.

7.5.5 What factors are associated with review issues?

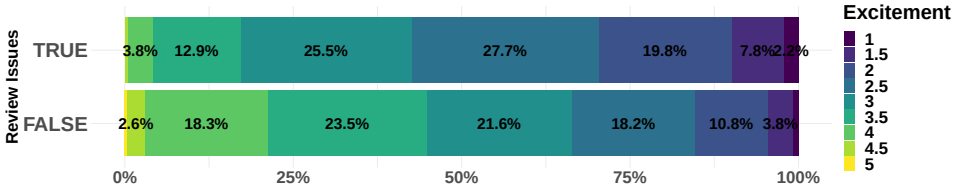
As discussed in §5.3, we introduced a mechanism for the authors to flag specific types of issues with reviews, and we received such flags for 12.9% of the reviews. Figure 11 shows the ratio of reviews with complaints (True) and without (False). For both *Soundness* and *Excitement* there is a clear trend towards more complaints with lower scores, but there are also complaints for high scores (e.g., 43.1% of reviews which the authors complained about had *Soundness* 4). This makes more sense if we consider the figure Figure 11d, which shows that 95% complaints are made about reviews where at least one of the scores is 3 or less. This suggests that reported review issues are associated with negative reviews, even for *Excitement* (although we tried to make it clear that this score is subjective and does not need arguing).

To explore other possible factors that could make the reviews more likely to be reported we fit a GLM model using the `glm()` function in R. The dependent variable is the presence or absence of reported issues (binary variable), and the predictors are the *Excitement* score (ordinal), *Soundness* score (ordinal), *Confidence* score (ordinal), *Reproducibility* score (ordinal), length of *Reasons to Reject* (interval), length of *Reasons to Accept* (interval), the *Contribution Match* (binary), *Area Match* (binary), *Language Match* (three-layer factor), *Reviewer's Experience* (three-layer factor), and *Reviewer's Activity* (binary). The link

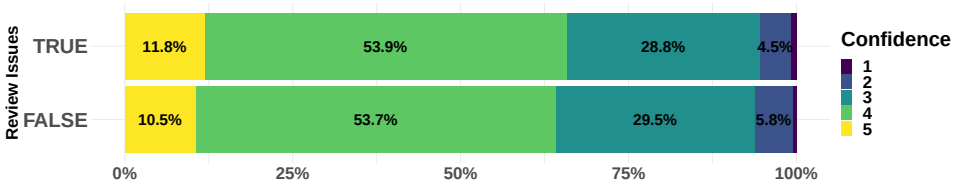
[32]This finding is important for the model reported in Table 4: the acceptance decisions are indeed based on both factors, and they are meant to capture different information, but the high correlation between these two variables suggests that the estimates obtained in Table 4 should be interpreted with caution.



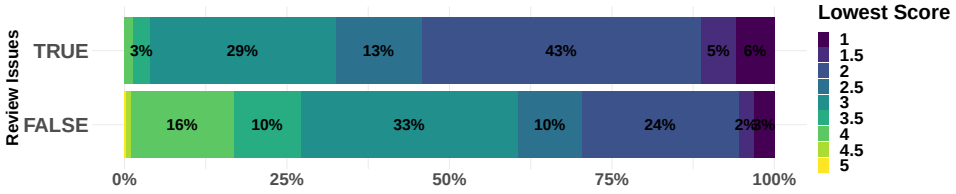
(a) Soundness



(b) Excitement



(c) Reviewer confidence



(d) Min(*Soundness*, *Excitement*)

Figure 11: Reviewer scores vs the amount of issues reported with reviews. In each plot, True/False refers to the reviews which were or were not flagged for review issues by the authors.

function was logit, corresponding to a binomial distribution of the response variable (logistic regression).[33] The coefficients of the fitted model are presented in Table 12.

We further employ the type III Anova using the ANOVA() function in R in order to obtain significance levels for each factor which are presented in Table 13. While McFadden's pseudo-R^2 of the fitted model is only 0.067, several variables of this model are significant predictors of the review issues.

The most significant factors are *Soundness*, *Excitement*, and the length of *Reasons to Accept*. All of these variables have a negative relationship with the reviewer issues, perhaps unsurprisingly, with higher scores the review is less likely to be reported. Similarly, longer text in the *Reason to Accept* field leads to less chance of the review being reported. Counter-intuitively, the positive coefficient associated with the reviewer being active suggests that when the reviewer is active (i.e. with at least one review revision or a forum message) the log-odds of the review issue increase by about 0.32, all else being equal. That is, the more active reviewers (putting in more effort) are actually receiving *more* complaints.

Other significant factors are *Language Match* and the reviewer's confidence; both associated with negative coefficients. This suggests that when the reviewer is familiar with the non-English language investigated in the study, the log-odds of a review issue decrease by approximately 0.26 (i.e., the review is 1.29 times less likely to be flagged for issues). Similarly, the negative coefficient of the reviewer's

[33] We inspect the residuals plots and compute the variance inflation factor to assure that the assumptions of GLM are not violated.

[34] Signif. codes: '$p < 0.001$' '***', '$p < 0.01$' '**', '$p < 0.05$' '*', '$p < 0.1$' '.', '$p > 0.1$' ' '.

[35] Signif. codes: '$p < 0.001$' '***', '$p < 0.01$' '**', '$p < 0.05$' '*', '$p < 0.1$' '.', '$p > 0.1$' ' '.

	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	0.5999	0.2570	2.334	0.0196	*
Soundness	-0.3816	0.0479	-7.967	1.63e-15	***
Excitement	-0.4584	0.0549	-8.349	< 2e-16	***
Confidence	-0.0855	0.0393	-2.176	0.0295	*
Reproducibility	0.0609	0.0335	1.816	0.0693	.
Reasons to Reject	0.0004	0.0002	1.508	0.1315	
Reasons to Accept	-0.0052	0.0011	-4.748	2.06e-06	***
Match Contribution (True)	0.0763	0.1148	0.664	0.5066	
Match Area (True)	-0.1352	0.1030	-1.313	0.1892	
Match Language (False)	0.0270	0.1476	0.183	0.8550	
Match Language (True)	-0.2639	0.1275	-2.070	0.0384	*
Experience (Experienced)	-0.0744	0.0684	-1.087	0.2769	
Experience (Zero)	-0.0274	0.1164	-0.235	0.8143	
Reviewer Active (True)	0.3172	0.0737	4.303	1.69e-05	***

Table 12: Coefficients of the Generalized Linear Model predicting the review issues. The table includes the coefficient estimate, standard error, z-value, and p-value for each predictor.[34]

	LR Chisq	Df	Pr(>Chisq)	
Soundness	64.65	1	0.0000	***
Excitement	70.45	1	0.0000	***
Confidence	4.71	1	0.0300	*
Reproducibility	3.31	1	0.0688	.
Reasons to Reject	2.23	1	0.1353	
Reasons to Accept	24.17	1	0.0000	***
Match Contribution	0.45	1	0.5035	
Match Area	1.69	1	0.1940	
Match Language	4.61	2	0.0998	.
Experience	1.24	2	0.5386	
Reviewer Active	19.35	1	0.0000	***

Table 13: Type III Analysis of Deviance for the variables in the Generalized Linear Model predicting whether issues were reported for the given review.[35]

Confidence suggests that with an increased *Confidence* score the likelihood of the review to be reported decreases though by a small margin.

7.5.6 Do we have bad actors?

To explore the possibility that many reported review issues are due to individual unprofessional reviewers, let us consider the fact that 1,620 reviews with reported issues were authored by 1311 reviewers, i.e. about a third of our total pool. But most of these reviewers had more than three reviews, and 1060 of them were only reported once. Of the remaining reviewers, 201 were flagged twice, and 50 reviewers had more than 3 complaints. We conclude that while there are indeed some unprofessional reviewers, and conferences need to systematically share such information and develop a system to address this problem, there are few such cases (6.2% if we consider all reviewers with more than 2 flags, and 1.2% with more than 3 flags). An interesting takeaway from Figure 11c is that the reviews that are problematic according to the authors, do *not* have lower confidence scores, so these are unlikely to be the new reviewers or the reviewers unfamiliar with the area.

According to folk wisdom, the bad reviewer is usually Reviewer2 (sometimes Reviewer3). We clear their good name: at ACL'23, the most issues were reported for Reviewer1, as shown in Figure 12.

7.5.7 Can the reviewers tell who the authors are?

In 567/12606 (4.5%) reviews the reviewers indicated that they have seen the paper, either by seeing a preprint (533) or by other means (34). Additionally, 513 (4.1%) reviewers indicated that they had a good guess of the author identity based on the paper content. 11460 (90.9%) ACL'23 reviews were reported as fully anonymous.

The community "recall" on the preprinted submissions is as follows: we had 628 submissions (13.8%

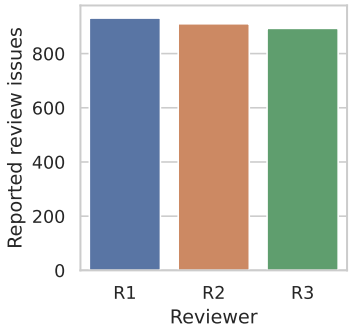


Figure 12: The number of review issues reported for R1, R2, and R3

of all direct submissions) for which the authors had disclosed preprints. The reviewers identified 306 (49%) of them. Hence, we estimate that although in our sample the number of "guesstimates" based on content is about the same as the number of preprinted papers, if the current 1-month embargo period was to be lifted, and the volume of preprints were to increase – the latter would also increase, while the volume of "guessed" authorship cases should stay the same (at about 4-5%). Interestingly, our reviewers reported another 102 submissions, for which preprints were not disclosed by the authors. We recommend that the future chairs investigate at earlier stages whether such cases are due to false memories of similar preprints, or preprint policy violations.

7.5.8 Do preprints affect the peer review process?

Having established that reviewers do have a high recall for preprints (§7.5.7), we investigate the possible connection between the reviewer's awareness of the author identity on their *Soundness*, *Excitement*, and *Confidence* scores by fitting Cumulative Link Mixed Effect models with the Laplace approximation using the `clmm()` function for the `ordinal` package in R (Christensen, 2022). The response variable is the given score and the predictor is the *Anonymity* answer (fixed effects). We also employ random intercepts for the paper (SubmissionID) and reviewer (ReviewerID) to account for this variability (random effects).[36]

Soundness. The results of the model fitted for the effect of *Anonymity* on the *Soundness* scores are present in Table 14. The *Anonymity* has five possible values: (1) the reviewer does not know the authors (reference level), (2) the reviewer may know the authors, (3) the reviewer knows the authors via means other than online posting, (4) the reviewer knows the authors via online posting prior to the anonymity period, and (5) the reviewer knows the authors via online posting post to the anonymity period. Estimates for different answers to the anonymity question presented in Table 14 suggest that the reviewers were 1.59 times more likely to assign higher *Soundness* scores when they thought they may know the authors, and 1.75 times more likely to assign higher *Soundness* scores when they have seen the preprint online.[37]

Excitement. The results of the model fitted for the effect of *Anonymity* on *Excitement* are present in Table 15. Estimates for different answers to the anonymity question presented in Table 15 suggest that the reviewers were 1.49 times more likely to assign higher *Excitement* scores when they thought they may know the authors, and 1.73 times more likely to assign higher *Excitement* scores when they have seen the preprint online.

Confidence. The results of the model fitted for the effect of *Anonymity* on reviewer's *Confidence* are present in Table 16. Estimates for different answers to the anonymity question presented in the table suggest that the reviewers were 1.29 times more likely to report higher *Confidence* scores when they

[36] We validate the model fit by examining residual plots and convergence criteria. The residual plots showed no clear patterns or extreme outliers, and the satisfactory convergence indicates a reasonable model fit. We further observe that, perhaps unsurprisingly, both SubmissionID and ReviewerID account for a substantial portion of the variability in each of the response variables.

[37] We take the exponential of each coefficient.

	Estimate	Std. Error	z-value	Pr(>|z|)	
Random effects:					
SubmissionID (Intercept)	2.2427	1.4976			
ReviewerID (Intercept)	0.7806	0.8835			
Fixed effects:					
Anonymity (2)	0.46037	0.11744	3.920	8.85e-05	***
Anonymity (3)	0.02567	0.41291	0.062	0.9500	
Anonymity (4)	0.55947	0.13081	4.277	1.90e-05	***
Anonymity (5)	0.36749	0.27565	1.333	0.1820	

Table 14: Cumulative Link Mixed Model Results for the effect of *Anonymity* on the *Soundness* scores. The reference level is Anonymity (1) (i.e., not knowing the authors).

	Estimate	Std. Error	z-value	Pr(>|z|)	
Random effects:					
SubmissionID (Intercept)	1.6675	1.2913			
ReviewerID (Intercept)	0.5163	0.7185			
Fixed effects:					
Anonymity (2)	0.39828	0.10629	3.747	0.000179	***
Anonymity (3)	0.13179	0.37724	0.349	0.726816	
Anonymity (4)	0.54498	0.11816	4.612	3.98e-06	***
Anonymity (5)	0.08329	0.24708	0.337	0.736049	

Table 15: Cumulative Link Mixed Model Results for the effect of *Anonymity* on the *Excitement* scores. The reference level is Anonymity (1) (i.e., not knowing the authors).

thought they may know the authors, and 1.80 times more likely to assign higher *Confidence* scores when they saw the preprinted online.

We thus conclude that submissions with preprints, as well as submissions where the reviewers believe they could guess the authors, systematically receive higher ratings for both *Soundness* and *Excitement*, as well as higher *Confidence* scores. We further note that preprinted papers are disproportionately recommended for consideration for best paper awards (and without such a recommendation from at least one reviewer the submissions are not considered by the best paper committee). In total, only 1.6% papers received any reviewer nominations at all, and for 30% of those papers, the authors had disclosed preprints.

While our data shows the pattern of higher scores, acceptance chances, and best paper nominations for preprinted submissions, the causal mechanism remains a question: is it because such papers are inherently higher quality, or because of the benefits of community feedback they may receive, or because of the well-documented reviewer biases towards towards famous names and institutions (Peters and Ceci, 1982; Tomkins et al., 2017, among many others)? Since these possibilities necessitate different actions on the part of the chairs who strive for higher-quality program, the causal question needs to be answered for informed policy decisions. Since we observe an increase in likelihood of higher scores both for real preprints and for submissions where the reviewers only thought that they might know the authors (although the effect is smaller in that case), we can conclude that the social factor is definitely present—but more research is needed to establish its exact contribution. But the fact that we only had 13.8% preprints suggests that the current 1-month embargo policy is effective in at least reducing the volume of the problem.

8 Special Review Processes

8.1 Ethics review

Following the practice started at NAACL 2021, we formed an Ethics Committee (EC) dedicated to ethical issues. The review process was based on work in prior conferences and further developed by ARR and recommendations from the ACL ethics committee. Initially there were 235 technical reviews flagging 218 papers for ethics concerns, and the SACs narrowed down the list based on the guidelines developed by the ethics chairs) to 75 papers, 6 of which did not make it to the ethics review (either withdrawn or cleared).

	Estimate	Std. Error	z-value	Pr(>|z|)	
Random effects:					
SubmissionID (Intercept)	0.416	0.645			
ReviewerID (Intercept)	3.413	1.847			
Fixed effects:					
Anonymity (2)	0.2576	0.1227	2.099	0.0358	*
Anonymity (3)	0.4210	0.4194	1.004	0.3155	
Anonymity (4)	0.5874	0.1342	4.376	1.21e-05	***
Anonymity (5)	0.3413	0.2864	1.192	0.2334	

Table 16: Cumulative Link Mixed Model Results for the effect of *Anonymity* on the *Confidence* scores. The reference level is Anonymity (1) (i.e., not knowing the authors).

20 papers under ethics review were labeled accept as-is, 43 received conditional accepts, and 6 were recommend for rejection. Of those recommended for rejection, 1 was accepted nonetheless, 1 was rejected as a result, and 4 were rejected on technical grounds. Of the conditionally accepted ones, 26 were rejected on technical grounds, and 1 was withdrawn. 16 passed the technical review and were conditionally accepted, meaning the ethics issues had to be addressed in the camera-ready version, to be verified by the SAC (based on EC guidance) prior to final acceptance.

The authors of all conditionally accepted papers submitted the camera-ready version and a short response that explained how they had made the changes requested. The SAC double-checked these revised submissions and responses, and confirmed that the ethical concerns had been addressed. As a result, all conditionally accepted papers were accepted to the main conference or Findings.

8.2 Best paper selection

ACL'23 implemented the new ACL award policy, aiming to expand the pool of work that is recognized as outstanding. In total, only 73 papers, i.e. 1.6% of all direct[38] submissions were nominated by the reviewers or ACs for consideration for awards. These papers were assessed by the Best Paper Award Committee, and with their help we selected 4 best papers, 4 special awards (social impact, resource, reproduction, theme paper), and 39 outstanding papers. The best and outstanding papers will be announced in a dedicated plenary session for Best Paper Awards on July 10 2023.

We encountered several issues with implementing the best paper policy as described in the wiki. With 73 nominated papers, to keep it down to 10 papers per judge and have 2 reviews per paper, we had to recruit 15 judges. At this scale, the workload is compatible with organizing a separate track: recruitment, paper assignments, chasing late reviews – only this time recruiting exclusively very senior and busy people, and it is very important to upheld diversity considerations (which we weren't able to do full justice). For the future, we recommend that a separate chair role is created for managing this process, similar in scope to the role of the ethics review chairs.

Furthermore, since the diversity considerations in the committee selection entail incompatible time zones, we found it impractical to require the judges to meet and jointly decide on the cases where they disagree (as recommended in the policy). Hence, after the judges cast their votes[39], the PCs made the final decisions on the basis of their recommendations (in particular, in the cases where one judge recommended outstanding paper and the other recommended not considering it further), we upheld the objections to flaws in the papers, shallowness of analysis, and ethical issues, which left us with 39 papers (a little short of the 1-1.5% total submissions policy target for the outstanding papers).

Finally, the ACL award policy described an Area Chair Award: the award that the SACs of a given track can give to one paper in their track, fully on their own authority. This was part of the guidelines for the final SAC recommendations, but we did not require them to be made at the same time. We sent out reminders after that, but received such nominations from only 12/26 tracks (with the theme

[38]This is only for the direct submissions to ACL. Due to the difficulty of seeing ARR nominations in START, we did not notice the 2 nominations out of 305 ARR submissions until it was too late.

[39]We found the agreement on the best paper committee votes to also be not very high: only 24/73 nominated papers received a unanimous vote to either consider for (any) award or not consider further.

track nomination transformed into the special Theme paper award). We recommend batching these recommendations with the final SAC recommendations as a single task.

9 Improving the Incentives

9.1 Improving Reporting Incentives for the Authors: Responsible NLP checklist

Following the effort started by NAACL 2022 and continuted at ACL Rolling Review (Carpuat et al., 2021), we used the Responsible NLP Checklist as a way to ensure that all submissions conform to a certain minimum standard of reporting on their reproducibility efforts, data collection principles, and consideration of broader impacts. However, at NAACL 2022 and ACL Rolling Review, these checklists are only used internally during peer review.

To improve the transparency of NLP research and create a stronger incentive to invest effort in this work, we made the Responsible NLP Checklists an official part of all published papers. The authors filled out the checklist information in a special form, and we later used that form to generate pdf versions of the checklist, which was appended to every paper pdf for the ACL Anthology.

This change was announced in our Call for Papers, and we additionally communicated it to the authors. The authors had the opportunity to update the checklist form during the preparation of the camera-ready version of their papers.

One modification to the checklist was introducing a mandatory question about AI writing assistance. This was motivated by the introduction of OpenAI's ChatGPT (OpenAI, 2022), the precedent of AI-assisted scientific paper writing of Meta's Galactica (Taylor et al., 2022), and, more importantly, a massive wave of promotion for AI "writing assistants" shortly before our direct submission deadline. We did not aim to completely ban AI-assisted writing (which does have legitimate use cases such as assistance to non-native English speakers), but to improve transparency: just like with the other ethics-related questions in the checklists, our posted policy required authors to explicitly state what they did. Our question and policy were subsequently adopted by ACL Rolling Review.

Magnusson et al. (2023) have reported that the higher rate of "yes" responses to the Reproducibility checklist at 4 NLP conferences. Given that our checklist includes reproducibility questions, and reproducibility positively correlates with both *Soundness* and *Excitement*, we would expect the Responsible NLP checklist to perform the same role. The reviewers themselves were predominantly positive about it: 66.99% rated it as "somewhat useful", 18.13% as "very useful", and only 14.35% — as "not useful".

Table 17 shows the ratios of submissions answering 'yes' to the questions of the checklist, and the acceptance rates for the submissions that answered 'yes' vs those that didn't. For most questions of the checklist, there is a small increase in acceptance rate for submissions that answer 'yes'. The most significant increases are for reporting limitations (so we recommend that the conferences keep mandating this section), reporting hyperparameters and computation budget (in line with the high correlation between reproducibility ratings and reviewer scores §7.5), citing relevant work, contributing scientific artifacts such as models and software (in line with our finding of a significant effect for this contribution type discussed in §7.3).

An interesting case is the "catch question" A3 (does your abstract accurately summarize your work?). It drew some criticism as "meaningless bureaucracy", since all submissions should respond "yes" to it. It was actually intended to see that the responders were not just clicking through the checklist. Most authors did respond 'yes', but those 2.24% that didn't saw a -25.4 decrease in acceptance rate. We interpret this as suggesting that the sloppiness in filling out the checklist correlates with sloppiness elsewhere in the work.

Finally, our new question about the use of writing assistants is the only one where the response 'Yes' is associated with a *decrease* in acceptance rate, although not very large.

9.2 Improving Incentives for Reviewers: Reviewer Awards

Arguably the biggest source of issues with peer review quality is the lack of incentives to invest more work in invisible service labor. One direction is *reputational* awards, eg via creating reviewer profiles, as in Publons. Another is *material* awards, such as monetary prizes similar to the best paper awards. Yet

Checklist question	*% submissions*	*Yes*	*Not Yes**	*Yes-Not_yes*
A1 (limitations)	46.92	47.62	17.05	**30.57**
A2 (risks)	56.23	49.28	43.88	5.4
A3 (catch question)	97.76	47.49	22.09	**25.4**
A4 (AI-assisted writing)	7.3	41.28	47.36	-6.08
B (artifacts)	72.45	50.09	38.58	**11.51**
B1 (cite)	71.02	49.96	39.46	**10.5**
B2 (license)	37.8	52.48	43.54	8.94
B3 (intended use)	45.28	49.48	44.8	4.68
B4 (PII)	22.02	49	46.33	2.67
B5 (documentation)	48.95	50.93	43.08	7.85
B6 (statistics)	70.47	49.76	40.14	9.62
C (computation)	92.31	47.76	36.82	**10.94**
C1 (parameters)	78.58	48.96	39.44	9.52
C2 (hyperparams)	85.5	48.49	37.63	**10.86**
C3 (stats)	81.02	48.19	41.51	6.68
C4 (packages)	76.01	47.16	46.15	1.01
D (humans)	28.98	52.11	44.8	7.31
D1 (instructions)	20.95	53.85	45.08	8.77
D2 (payment)	21.19	53.5	45.15	8.35
D3 (consent)	17.31	51.2	46.02	5.18
D4 (IRB)	9.62	53.24	46.25	6.99
D5 (demographics)	14.61	54.27	45.66	8.61

Table 17: The ratio of 'Yes' responses to checklist questions vs the responses other than 'yes' (i.e. both 'no' and 'no response'). The average acceptance rate in this pool is 46.92%.

another is *punitive* incentives, such as penalizing the late reviewers by delaying the reviews for their own submissions (Hauser and Fehr, 2007), or even blocking them from reviewing at future conferences.

All of these approaches are not without issues. Punitive incentives generally shift the focus to not getting penalized, rather than delivering high-quality reviews. Material awards may introduce the wrong incentives (Squazzoni et al., 2013), and, depending on the institution and the country, the prize may be taxed or not even make it to the recipient. Conference fee waivers also may also reward the reviewer's institution rather than the reviewer, since the institutions usually bear the registration costs. While a survey found that reviewers generally prefer reputational awards over material (Warne, 2016), their value also depends on whether the reviewer's institution rewards such work.

We proposed to the ACL exec (and received their approval for) an initiative to match the new ACL best paper award policy with recognizing about 1-1.5% of outstanding reviewers and chairs. This combines reputational and material incentives. Instead of monetary prizes, we proposed awarding vouchers for virtual attendance of any *ACL (ACL, NAACL, EACL, AACL, EMNLP) conference of the awardee's choice, to be used within a year of the award date. Since many institutions do not support the attendance of conferences without accepted papers (or even with papers accepted to workshops and Findings), we hope that this measure will increase the overall number of conferences that the awardees can attend.

We asked the area chairs to nominate the reviewers in their pool who provided extra helpful reviews, high-quality emergency reviews, "champion" reviews, reviewers who were particularly active in the discussion phase, or demonstrated exceptional open-mindedness or expertise. We received 51 such nominations. We also asked the Senior Area chairs to nominate exceptional area chairs, receiving 13 nominations. Finally, we as the program chairs also nominated the (3) SACs of the track who were the most on-time, provided the most helpful feedback, and followed our instructions the most closely. Excluding the duplicates, this resulted in 67 total nominations. All awards will be announced on the conference website[40].

Since the total number of nominations was within our target number of awards (1-1.5% of total reviewers and chairs), we were able to award all 66 nominations (out of 4998) without creating a selection committee. In the future, we recommend that an extra volunteer role is created for managing the selection of awardees and managing the awards.

[40] https://2023.aclweb.org/program/best_reviewers

Caveats: despite our calls to nominate reviewers and chairs, relatively few ACs and SACs did that: only 7/70 SACs and 28/438 ACs. We recommend that the AC/SAC guidelines are expanded with a section about these awards, and that ACs are asked to start keeping track of potential outstanding reviewers at the (a) review quality check stage, (b) discussion stage, rather than only during meta-reviews (as we did). The SACs could be asked to start keeping track of outstanding ACs at the (a) assignment checks, if that is the process used by the venue, (b) meta-reviews, (c) nominating on the basis of quantitative analysis of the activity in the discussion forum and the number of author-reported review issues that the AC addressed.

9.3 Improving Incentives for Chairs: Peer Review Reports

Our final proposal for improving the incentives for peer review work was to increase its visibility by placing the program chair reports and any findings from their analysis of the internal conference data as an official part of the proceedings for the respective conference. This report is aiming to create a precedent for that. In the past, there have been two options for publishing such work: standalone research papers that undergo their own peer review, and miscellaneous blog posts and reports published in ACL wiki. But the former is not appropriate for reporting on incidental findings (since most of the program chairs work is not executed as a research project targeting a specific research question). The latter is unfortunately too difficult to discover, especially for the people outside of our field or new organizers who may not know which blog posts and wikis to search.

This initiative aims to improve the transparency of the overall process, and lets the younger members of the community have more insight into how the *ACL conferences work. Moreover, given the increasing attention to peer review in NLP community (Gao et al., 2019; Caragea et al., 2019) and more broadly in ML conferences (Price, 2014; Stelmakh, 2020; Beygelzimer et al., 2021), it would be useful to make the incidental findings from the conferences more easily discoverable, incl. to the researchers in the ML community and other fields.

The main difficulty for the program chairs and the publication chairs with implementing this proposal is that the full report needs to be prepared before the conference, when there is a lot of other work. To implement this, the set of volunteer roles would need to be expanded (see section 10). We also recommend that to the extent possible, the future chairs start documenting their workflow for the report early on (perhaps during the main review cycle).

10 Recommendations

Improving logistics. There are several sources of papers to the ACL main conference that the program chairs have no control over: TACL, CL, Industry Track Papers, SRW papers. This means that the PCs need to ingest four different sources of information with potentially little means of interacting with the relevant authors (in contrast to direct submissions). ARR is in a liminal space between direct submissions and these other papers. The timing and format of how the papers enter ACL should be standardized.

Desk rejections. Desk reject requirements should be clearly stated in the call for papers or in the ACL Paper formatting guidelines. The guidelines omit rules or lack clear thresholds for rejection. For example, there is no minimum separation between captions and tables/figures nor between section titles and the text above and below. Nor are there minimum text sizes for text within tables or figures. Adding clear rules would make the first pass reviewing more efficient and fair. ACL also needs to communicate more clearly about the role of the `aclpubcheck` script: it's a necessary but not sufficient check. Many authors assume that if they pass the `aclpubcheck` script, then they have followed all formatting guidelines.

***Soundness/Excitement* scores.** With predominantly positive feedback in the exit survey (§6.4), and evidence of significant improvement in raw agreement (§7.4), we believe this experiment was successful and should be continued. The formulation of the scores and the review form should be improved, and care should be taken to reduce the overall complexity of the form.

Review issue flagging. This feature received overwhelming support from the authors, and should be continued and standardized (i.e., cleanly incorporated into author response form)—especially since it

is likely to improve after several iterations, when everybody is more familiar with it and the reviewer guidelines. More AC training is needed to address the flagged issues.

Continued reviewer policy publications. 12.9% of all ACL'23 reviews were flagged by the authors for various issues, with the most frequent problem being reviewer heuristics such as "too simple" and "not SOTA". It is reassuring to know that the ratio of bad reviews is already not very high, but of course we should strive to further decrease it. The reviewer guidelines, in combination with the review issue flagging mechanism, serve a double purpose: even if the reviewers do not read them, the authors will (since they have the incentive to call out problematic reviews), and then the area chairs also will (to handle the author-flagged issues). Hence, eventually, these policies will become widely known across the community, and enforced by it. We urge the future chairs to continue publishing their reviewer policy or simply re-use ours, and explicitly point to it in review, author response, and meta-review forms.

Reviewer assignment check support. There is currently no convenient interface for the ACs to look up the assigned reviewers and browse the alternatives with up-to-date availability information. Its lack is a major hurdle for the chairs, and it may cause either delays in the process or skipping the checks.

Reviewer match explanations. Our area chairs were very positive about this feature. For venues not using an interpretable assignment algorithm such as our keywords-based process, at the very least, the reviewer profiles and relevant papers should be provided directly with the review, without any extra search.

Post-acceptance decision litigation. Having increased the acceptance rate for Findings, we were surprised to still receive a large volume of emails from the authors who, considering their scores and meta-review, argued that either their paper should have been accepted to the main track, or that it shouldn't have been rejected. It appears that some subcommunities share their scores with each other, under the mistaken impression that if one paper with certain scores was accepted, others with similar scores should be too. We had no capacity for anything beyond checking for clerical errors. The peer review process is by no means perfect, and there was certainly some noise in the decisions—but it is also certain that many authors who disagree with their decisions would try to argue their case if given the chance. If such litigation is not an announced an official part of the conference process—doing so for the select few would not be fair to all the other authors who also disagree with their decisions. We recommend that the future chairs either build this into their process and dedicate time and resources to it, or pre-announce that decisions are final and will not be reconsidered, beyond the cases of clerical errors.

Area-Contribution-Language matching. The results of our experiment with exactly matching the reviewers with submissions by these areas allowed us to establish that it is possible to ensure a fair acceptance rate for most "non-mainstream" contribution types, and for the 63.8% of the submissions that had target languages other than English, we were able to provide a reviewer competent in that language. These results are by no means perfect, and it is important that the future venues improve on them, perhaps with other methods. But Area-Contribution-Language matching could be considered a fair baseline for the future conferences, when considering the success rates for different types of submissions and languages. All that is needed from the chairs is to include in submission forms the checkboxes for different types of contributions, and input fields for the target languages other than English. At the very minimum, the chairs would then be able to analyze the acceptance rates of different types of submissions, and compare it with ours (Table 7). One step further would be to also solicit this information from the reviewers, and estimate the quality of automated matches by the explicit keyword matches (see Table 2).

One more practical takeaway for future work is that if we used a solution relying purely on publication history from Semantic Scholar—25% of our matches would have been made on unreliable information. For embeddings-based solutions to work better, we would first need to provide them with better data, and this will take a bigger Semantic Scholar cleaning campaign than what we were able to elicit.

Reconsidering the acceptance rate for Findings. The initial iterations of Findings starting with EMNLP 2020 had the Findings acceptance rate at about 35%. This is the target rate we gave to our SACs, and then we tried to accommodate as many of their ranked preferences as we could. Although

we had over 40% rate with Findings, still, in many SAC comments we saw that they were overriding acceptance recommendations of ACs only to meet the quotas. While the quota for the Main track will stay at 20-25% for venue ranking reasons, we do not see why Findings could not be further extended to have room for most sufficiently sound work. About 60% of our direct submissions had at least two positive (above-borderline) reviews for *Soundness* and at least one for *Excitement*. Assuming some noise in the negative reviews for *Soundness*, it would be only reasonable to expect that at least 45%-50% submissions are Findings-worthy. Of course, the track SACs would not *have* to accept that many (the ratio of high-quality papers may vary between tracks and years), but when they do not see good reasons to reject — they should not be constrained by the Findings quota. This step would presumably also further decrease the burden of re-reviewing for resubmissions. We also recommend developing a standard process for Findings authors to apply for presentation at topically matching workshops, and for at least virtual poster presentation slots at the main conference.

Further research on the effect of preprinting on peer review. We find that the preprinted papers have consistently higher ratings (for both *Soundness*, *Excitement*, and reviewer confidence), get more recommendations for awards, and a higher acceptance rate. There are several possible underlying causes (from reviewer biases to higher initial paper quality and benefits of community feedback), which likely all contribute to this effect. Since these factors necessitate different actions if they were the major contributor to the observed effect, for informed policy decisions it is necessary to establish how they intermix. We observe however that although the present 1-month embargo policy does not solve this problem, it is effective at mitigating it, since we only had 13.8% such papers.

Consistently working to improve peer review concistency. Our analysis shows that the inconsistency in numerical reviewer score ratings is remarkably consistent across *ACL conferences (at about α 0.3 across EMNLP'22, EACL'23, and ACL'23). Among the likely culprits are miscalibrated scales, different interpretations of scales, at least some reviewers not even reading the guidelines, and reviewer biases. That said, we do see almost twice the raw agreement for our *Soundness* score (that is supposed to be more objective) over *Excitement* (more subjective), when the scores are mapped to the sound/unsound vs exciting/unexciting categorical variables. This suggests that asking more concrete questions does help (as long as the reviewer form does not become too complicated), and we can continue improving peer review on the basis of the general NLP methodology for iterating on guidelines and measuring agreement.

Ethics review. The innovation of the ethics review is useful and necessary, but it should be explicitly built into the timeline. We particularly struggled with the conditional accepts.

Responsible NLP Checklist. With predominantly positive reviewer feedback and evidence of improved acceptance rates for submissions that follow the best reporting practices, we believe that this is an important instrument for creating the right incentives for better science. We also recommend continuing to make it public, to strengthen these incentives.

AI-assisted reviews. We did not expect this happen so soon, but already at ACL'23 some chairs reached out to us with questions about reviews that they suspected to be at least partly generated. The reviewer guidelines will need to be updated with respect to that as well, including how sending papers to cloud-based language models may violate confidentiality.

Review policy updates. The rise of popular commercial systems such as ChatGPT that are claimed to be general-purpose, made an unfortunate match with our field's tendency to expect the popular systems in all papers as universal baselines. We did not consider this at ACL'23, since ChatGPT fell out of scope of 3-month policy for considering contemporaneous work, but we did already have at least one precedent of a reviewer asking for a comparison with ChatGPT. We recommend that future chairs develop a clear policy in the reviewer guidelines about requests for comparisons with "closed" systems, to avoid numerous issues with evaluation methodology and benchmark data contamination (Rogers, 2023).

Expanding the set of volunteer roles. Our experience suggests that PC-ing a conference of ACL'23 size is a job that can no longer be realistically done by 3 volunteers. Early on, we introduced a ***visa***

support team[41] to start early with issuing the letters of invitation for Canada. We also had crucial help from two ***PC assistants***: Youmi Ma, an administrative assistant who handled much of the conference email, and Marzena Karpinska, who helped with analysis of peer review data in this report. In the future, we recommend that a dedicated role of a ***peer review chair*** is created, whose responsibility will be to supplement PC report with analysis of the data of the respective conference and comparing it with any records from previous conferences (so as to establish the effect of any new policies), and to coordinate the peer review awards selection and logistics (see §9.2). The growing volume of nominations for best papers requires a ***best paper chair***, handling in effect the organization of a separate track and review process. Finally, we could have used a lot of help in the conference schedule: ideally there would be a dedicated ***schedule chair***, ideally serving at several conferences so as to reduce friction and reuse the skill set as much as possible, as well as incorporate feedback from several events. Given that ACL had papers from SRW, Industry, ARR, TACL, CL, Findings, and the Main Conference, it's not necessarily feasible that the main track PCs can effectively coordinate scheduling all of these papers.

Another option would be for each conference to have ***two sets of PC chairs, one remaining from the previous year and one new***. This would lighten the workload and ensure a smoother process (since people do not learn how to do everything from scratch each time). The first-year PCs would do the bulk of the work after the paper notifications are sent, and the second-year PCs would concentrate on the review process, analysis and the report. The first-year PCs would observe that and have better knowledge for designing the review process (CFP, SAC nominations, review criteria, etc).The second-year PCs would observe the COI requirements.

11 Acknowledgements

ACL'23 was the result of an incredible effort of 70 SACs, 438 ACs, 4490 reviewers, and 13,658 authors. We also thank our 2 ethics chairs and their 21 reviewers, as well as 15 judges on the best paper committee.

We thank the ARR team, and particularly Jonathan K. Kummerfeld, Thamar Solorio and Mausam, for their help with integrating ARR submissions and analyzing them.

We had a chance to learn from the past chairs Smaranda Muresan, Preslav Nakov and Aline Villavicencio (ACL 2022), Yoav Goldberg, Zornitsa Kozareva, Yue Zhang (EMNLP 2022), and Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur (NAACL 2021). We also thank EMNLP 2022 and EACL 2023 (Isabelle Augenstein, Andreas Vlachos) for sharing their score distribution data for our analysis.

Our work is built on many iterations of previous *ACL conferences, including the AC and SAC guidelines developed at ACL 2021, and peer review tutorials developed by Anna Rogers and Isabelle Augenstein for ACL Rolling Review.

Our paper-reviewer matching relied on Semantic Scholar data, kindly provided by Kyle Lo (AI2). The Semantic Scholar team also provided extra support to numerous authors working to clean up their profiles.

Emma Strubell, Ian Magnusson, and Jesse Dodge helped us to prepare publishable versions of Responsible NLP checklist.

We were only able to devote that much effort to peer review and its analysis thanks to the help of our brilliant assistants Youmi Ma and Marzena Karpinska.

Richard Gerber (START) responded to numerous issues and implemented several changes at our request, including the possibility to include "explanations" for the paper-reviewer matching.

We deeply thank the ACL Executive (especially Iryna Gurevych, Tim Baldwin, David Yarowsky, Yusuke Miyao, and Emily M. Bender) for their support of many of our crazy ideas, including the reviewer awards and the publication of this report.

Last but not least, we thank our publication chairs and ACL Anthology team, in particular, Ryan Cotterell and Matt Post — for their infinite patience with this last-minute publication.

[41] https://2023.aclweb.org/blog/visa-info/

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif M. Mohammad, and Karën Fort. 2023. The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research.

Omer Anjum, Hongyu Gong, Suma Bhat, Wen-Mei Hwu, and JinJun Xiong. 2019. PaRe: A Paper-Reviewer Matching Approach Using a Common Topic Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 518–528, Hong Kong, China. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Rachel Bawden. 2019. One paper, nine reviews.

Emily M. Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters.

Emily M. Bender and Leon Derczynski. 2018. Paper Types.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. The NeurIPS 2021 Consistency Experiment.

Cornelia Caragea, Ana Uban, and Liviu P. Dinu. 2019. The Myth of Double-Blind Review Revisited: ACL vs. EMNLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2317–2327, Hong Kong, China. Association for Computational Linguistics.

Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. 2021. Responsible NLP research Checklist.

Rune Haubo Bojesen Christensen. 2022. ordinal—Regression Models for Ordinal Data. R package version 2022.11-16.

Kenneth Ward Church. 2020. Emerging trends: Reviewing the reviewers (again). *Natural Language Engineering*, 26(2):245–257.

Trevor Cohn, Yulan He, Yang Liu, and Bonnie Webber. 2020. Advice on Reviewing for EMNLP.

Corinna Cortes and Neil D. Lawrence. 2021. Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. *arXiv:2109.09774 [cs]*.

D.R. Cox and E.J. Snell. 1989. *Analysis of Binary Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-Yes-Yes: Proactive Data Collection for ACL Rolling Review and Beyond.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Marc Hauser and Ernst Fehr. 2007. An Incentive Solution to the Peer Review Problem. *PLOS Biology*, 5(4):e107.

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. Association for Computational Linguistics.

Letizia Jaccheri, Cristina Pereira, and Swetlana Fast. 2020. Gender Issues in Computer Science: Lessons Learnt and Reflections for the Future. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 9–16.

Steven Jecmen, Minji Yoon, Vincent Conitzer, Nihar B. Shah, and Fei Fang. 2022. A Dataset on Malicious Paper Bidding in Peer Review.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

Andy Liaw and Matthew Wiener. 2002. Classification and Regression by randomForest. *R News*, 2(3):18–22.

Michael L. Littman. 2021. Collusion Rings Threaten the Integrity of Computer Science Research. *Communications of the ACM*, 64(6):43–44.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What Have We Learned from the Checklist?

Daniel McFadden. 1973. Conditional Logit Analysis of Qualitative Choice Behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press New York, New York, NY, USA.

Nico Nagelkerke. 1991. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

OpenAI. 2022. Introducing ChatGPT.

Katarina Pantic and Jody Clarke-Midura. 2019. Factors That Influence Retention of Women in the Computer Science Major: A Systematic Literature Review. *Journal of Women and Minorities in Science and Engineering*, 25(2).

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*. Springer International Publishing.

Douglas P. Peters and Stephen J. Ceci. 1982. The Fate of Published Articles, Submitted Again. *Behavioral and Brain Sciences*, 5(2):199–199.

Eric Price. 2014. The NIPS experiment.

Anna Rogers. 2023. Closed AI Models Make Bad Baselines.

Anna Rogers and Isabelle Augenstein. 2020. What Can We Do to Improve Peer Review in NLP? In *Findings of EMNLP*, pages 1256–1262, Online. Association for Computational Linguistics.

Richard Smith. 2010. Classical Peer Review: An Empty Gun. *Breast Cancer Research*, 12(4):S13.

Charles Spearman. 1987. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 100(3/4):441.

Flaminio Squazzoni, Giangiacomo Bravo, and Károly Takács. 2013. Does Incentive Provision Increase the Quality of Peer Review? An Experimental Study. *Research Policy*, 42(1):287–294.

Ivan Stelmakh. 2020. Experiments with the ICML 2020 Peer-Review Process.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pages 828–856. PMLR.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science.

Terne Thorn Jakobsen and Anna Rogers. 2022. What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4810–4823, Seattle, United States. Association for Computational Linguistics.

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.

William N. Venables and Brian D. Ripley. 2002. *Modern Applied Statistics with S*, fourth edition. Springer, New York. ISBN 0-387-95457-0.

Verity Warne. 2016. Rewarding reviewers – Sense or Sensibility? A Wiley Study Explained. *Learned Publishing*, 29(1):41–50.