ARR-Dv2: New Data from ACL Rolling Review and its Potential to Improve Reviewing in the ACL Community

Anonymous ACL submission

Abstract

Peer review is a central part of the scientific publication cycle, and the peer review service has long been an honor and obligation that researchers gladly fulfill. With the considerable increase in submissions, this aspect of voluntary contribution to the research community has turned into a burden due to the increased workload involved. We present a new dataset to *support* AI-assisted and human studies of the peer reviewing process, based on the ARR Open Review Platform. Unlike previous similar data collections, ours (ARR-Dv2) not only contains the reviews but also data from the author-reviewer discussions (aka rebuttal phase) and meta-reviews. We present statistical analyses of the data with respect to the overall score and correlation between the overall score, confidence, and soundness scores. Additionally, we extend the task of *Review Score Prediction* to also include rebuttal data and analyze its effect on score prediction. Finally, we introduce a new task: *Meta Review Score Prediction*, which is based on a set of up to three reviews rather than a single review. Our initial results in a zero-shot setting indicate that the rebuttal data adds valuable information to the score prediction and allows for a reliable prediction especially when considering a divergence of 0.5 points in the scores.

1 Introduction

Peer review is the key mechanism for quality control of scientific manuscripts across academia (Birukou et al., 2011), playing a vital role in ensuring research integrity and reliable progress (Bornmann, 2011). Commonly, a group of researchers (the *reviewers*) scrutinizes the manuscript and each composes a *review report* that serves as the basis for a *meta-reviewer* to create a condensed overview of the individual reviews, known as *meta-review report*. Despite its importance to the scientific community with the growing number of publications in

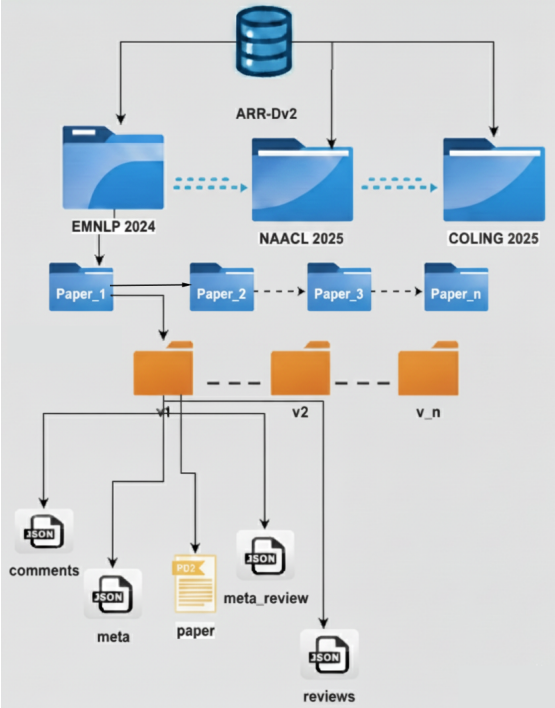


Figure 1: The structure of ARR-Dv2 dataset

science (Landhuis, 2016) and Artificial Intelligence (AI) research in particular (Sculley et al., 2018), the reviewing load places significant pressure on researchers. This amplifies existing reviewer biases (Lee et al., 2013), encourages outsourcing to junior researchers (Kuznetsov et al., 2024b), or – with the rise of generative AI (GenAI) tools – to Large Language Models (LLMs) such as ChatGPT (Achiam et al., 2023) and the like (Liang et al., 2024). This, in turn, introduces new biases and quality issues (Du et al., 2024). Putting the reliability of peer review at risk conceivably has severe consequences for the scientific discovery process as a whole (Ebadi et al., 2025; Hanafi et al., 2025).

Human-AI collaboration and assistance for peer reviewers is a promising tool to mitigate these issues by speeding up the review process without sacrificing quality, and while maintaining full con-

trol for human reviewers (Kuznetsov et al., 2024a). Training and evaluation data are essential for the development of such systems. While there is a plethora of peer reviewing corpora (Yuan et al., 2022; Zhang et al., 2025), most of them come from AI research and open reviewing systems. However, reviewing behavior varies between domains and reviewing systems, hereby limiting the generalization of findings to other domains (Kuznetsov et al., 2024b). Additionally, using publicly available reviewing data for evaluation poses risks of contamination (Balloccu et al., 2024; Jiang et al., 2024; Ravaut et al., 2024). In this work, we introduce ARR-Dv2, a new peer review corpus collected from ACL Rolling Review (ARR) – a closed reviewing system that ensures its data has not been seen during training of current LLMs – to support the development of peer review assistance systems tailored to reviewing in natural language processing (NLP) research. Figure 1 shows the structure of ARR-Dv2.

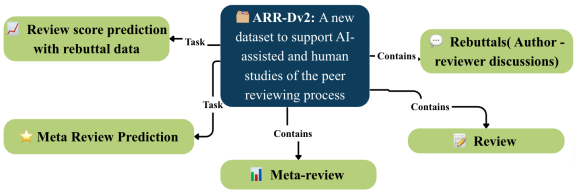


Figure 2: ARR-Dv2 Dataset

This new corpus complements previous work by Dycke et al. (2023) covering NLP reviewing data prior to 2023. However, NLPEER includes only review reports without author rebuttals and reviewer discussion or meta-reviews. ARR-Dv2 includes these important reviewing artifacts (see Figure 2) and is the largest and most recent peer review dataset from the ACL community so far, including the conference of Empirical Methods in Natural Language Processing (EMNLP) 2024, Nations of the Americas Chapter of the ACL (NAACL) 2025, and the International Conference on Computational Linguistics (COLING) 2025. We make the data easily accessible for further research. The dataset enables studying the submitted and accepted version of each paper, their reviews, and the reviewer-author discussion with meta-reviews with open licenses and ethical clearance.

Based on the newly collected ARR-Dv2, we quantify the reviewers' behavior across time and venues. We find that reviewing in NLP and AI across time does show a focus on similar aspects of papers, such as novelty and experimentation, but with varying importance. Additionally, we turn to a practical assistance task during peer review, *review score prediction*, aiming to support reviewers in choosing adequate scores for their textual report; this is particularly challenging for junior reviewers and reviewers from other fields, and may ease their adaptation to the NLP domain. We use the new reviewer-author discussion (rebuttals) and meta-review data from our ARR-Dv2, to investigate these artifacts' effect on score prediction performance. In summary, we contribute

- an extension to the ethically and legally aware collection protocol of Dycke et al. (2022),
- a new corpus of papers, reviews, meta-reviews and author-reviewer discussions from the NLP community from 2024-2025,
- a new perspective on rebuttal and meta-review data during review score prediction lying the groundwork for the systematic investigation of rebuttals' effect on reviewers' scoring, and
- the exploration of the task of meta-review score prediction in the ACL community.

2 Related Work

Peer Reviewing Corpora have been published over the past years. Most datasets originate from machine learning (ML) conferences such as ICLR and NEURIPS (Staudinger et al., 2024) crawled from open venues hosted on OpenReview[1] (Yuan et al., 2022; Kennard et al., 2022, i.a.) with the most recent and complete version presented by Zhang et al. (2025). Peer review data from other domains includes closed corpora, such as the Elsevier corpus (Willems, 2022), and open reviewing platforms, such as F1000Research[2], which are included in NLPEER (Dycke et al., 2023). Unlike most other datasets, ARR-Dv2 includes peer reviews from a closed-reviewing system, which is currently underrepresented despite its dominance in science (Dycke et al., 2023) and ensures contamination-free evaluation, as the reviews cannot have been part of the pretraining data of current LLMs yet. Only the PeerRead corpus (Kang et al., 2018) and the ARR and COLING subsets of NLPEER (Dycke et al., 2023) cover peer reviewing data from the NLP community. Both corpora do not include meta-reviews or author responses, and they cover substantially

[1] https://openreview.net
[2] https://f1000research.com/

fewer papers and reviews compared to ARR-Dv2. Table 1 summarizes the key differences between NLPeer and ARR-Dv2.

Review Score Prediction aims to infer the numerical review scores (e.g., Soundness, Excitement) typically accompanying review reports in the NLP community. These scores are critical to informing meta-reviewers and acceptance decisions. This task has gained increasing attention in recent years. Kang et al. (2018) are the first to study review score prediction, training a simple regression model. Ribeiro et al. (2021) investigate the feasibility of review score prediction based on sentiment labeling of review texts and SVMs. Bharti et al. (2021) are the first to propose a neural model that combines both the paper and review text to predict review scores. Dycke et al. (2023) evaluate BERT-based models for review score prediction across domains. Fernandes and Vaz-de Melo (2024) conduct a large-scale evaluation of review score prediction models and identify specific inputs that challenge current models. Finally, review score prediction accuracy also has become an important metric for assessing the quality of automatic review generation systems (Yuan and Liu, 2022; Lu et al., 2024; Zhou et al., 2024; Idahl and Ahmadi, 2025; Zhu et al., 2025; Weng et al., 2025). With the newly collected ARR-Dv2 dataset, it is now possible to extend this task to meta-review score prediction using data from the ACL community, whereas prior work has primarily relied on review data from other communities, most notably the machine learning community (Bharti et al., 2021; Lu et al., 2024; Zhu et al., 2025). Additionally, the inclusion of rebuttal data allows us to investigate how rebuttals influence reviewers' scoring, providing further insight into the dynamics of peer review.

3 ARR-Dv2

The following section presents details on the data collection process, the resulting dataset, and various initial analyses.

3.1 Data Collection

Peer reviewing in the ACL community is a highly confidential process, where great care is taken to keep the information about authors and reviewers secret. Dycke et al. (2022) presents a donation-based data collection protocol — called the Yes-Yes-Yes- or 3Y-workflow — designed for collecting data from sensitive domains with multiple stakeholders and applying it to peer reviews. Their protocol, developed in consultation with the ACL ethics committee, offers an approved framework for collecting and handling peer review data. Our work extends the existing protocol to include author-reviewer discussions and meta-reviews in the licensing workflow, thereby increasing coverage of the donated artifacts. All modifications have been approved and developed in coordination with the ACL Ethics Committee. Figure 3, summarizes the data collection procedure.

Figure 3: The procedure and checks we performed to ensure that the data collection was as ethically sound as possible.

Extended Protocol The 3Y-Workflow is a default opt-out three-step workflow that first asks reviewers to donate their reports, then checks for acceptance of papers, and finally reaches out to authors of accepted papers for their approval of data release. As the authors observe, the coverage of donated papers is low, which is arguably attributed to the long time period between submitting a paper and its acceptance, disincentivising authors to donate their review data and drafts. We modify the procedure, maintaining default opt-out and decoupling from the review process, to ask authors for a donation decision during submission with an option to revise this decision throughout the review process.

As a second major change, we extend the license transfer agreements of the 3Y-workflow to include the author-reviewer-discussion and rebuttals asking authors and reviewers for license transfer for these artifacts during donation. Analogous to reviewers, we ask meta-reviewers to donate their meta-reviews during the reviewing period conditioning their inclusion in the dataset on the authors explicit approval.

All steps involve explicit disclaimers for all stakeholders. The donation of data is voluntary and can be revoked at any time. See Appendix A for more details on the collection workflow.

Collection We collected ARR-Dv2 in the time period from 2024 to 2025. In accordance with the ACL Rolling Review board, we retrieved the re-

	ARR@NLPeer	ARR-Dv2
time	2021-2022	2024-2025
#papers	476	2838
% accepted	100	100
#reviews	684	1923
#papers w/ rebuttals	-	1403(52.5 %)
Avg. sent per rebuttal	-	15.0 ± 13.4
#reviews per paper	1.43 ± 1.1	0.68 ± 0.75
#sent per paper	220 ± 62.9	538.9 ±199.6
#sent per review	31 ± 27.0	22.9 ± 12.3
total #tok in reviews	266k	822k
total #tok in rebuttals	-	1.8M

Table 1: Comparison of NLPeer (Dycke et al., 2023) and ARR-Dv2in terms of papers, reviews, rebuttals, etc. We report the mean and standard deviation of per-paper and per-review statistics.

viewing data using the OpenReview API based on the codebase of Dycke et al. (2022). Compared to the earlier NLPeer dataset, which covers 476 ARR submissions and 684 reviews, ARR-Dv2 offers a substantial expansion: it includes 2838 papers and 1923 reviews, with 52.5% of papers accompanied by rebuttals.

3.2 Meta-data Analysis

We analyze the metadata of ARR-Dv2 collected from EMNLP 2024, COLING 2025, and NAACL 2025 to determine differences from the previously collected data from the ACL community.

Comparison to NLPeer As shown in Table 1, ARR-Dv2 covers more data compared to the ARR subset of NLPeer. The number of papers increased from 476 to 2838, and reviews grew nearly threefold from 684 to 1923. ARR-Dv2 includes rebuttals for more than half of the papers (52.5%), with an average of 15 sentences per rebuttal. Notably, the papers from 2024/25 are longer, while their reviews are shorter, compared to the data up to 2022. This may be because reviewers had more papers to review as the number of submissions increased, and also because the papers themselves often included longer appendices and reference lists.

Review Score Distribution Review scores are a useful proxy for reviewer sentiment. We estimate their distribution in the dataset to quantify the diversity of papers with respect to quality and reviews in terms of sentiment in ARR-Dv2.

Figure 4 shows the distribution of reviewers overall scores (1-5) for each venue in ARR-Dv2. We observe that the distribution of the score is more

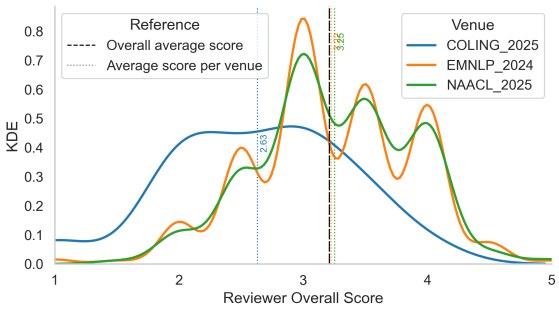


Figure 4: Distribution of review overall scores by venue. Kernel density estimates(KDE) curves show the score distribution for each venue. Colored dotted lines show the average review score per venue(COLING ≈ 2.63, EMNLP ≈ 3.23, NAACL ≈ 3.25), while the black dashed line indicates the overall average scores across all reviews.

shifted toward lower values for COLING 2025, whereas NAACL 2025 and EMNLP 2024 concentrate around 3 to 4, and the highest score (5) is rarely given to papers.

Meta-review Score Distribution A meta-review, usually written by the Area Chair (AC), summarizes the reviews and discussion for a submitted paper (Wu et al., 2022). In Figure 5, we look at their distribution in ARR-Dv2 to see the typical score, how spread out the scores are, and whether different venues use the scoring scale differently. The

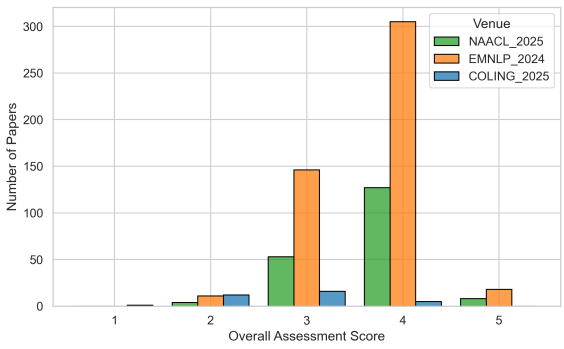


Figure 5: Number of papers by meta-review overall assessment score distribution for each venue.

meta-review scores for EMNLP 2024 and NAACL 2025 are primarily at 4, with some at 3, while COLING 2025 has more papers scoring between 2 and 3 and fewer at 4. We observe that some papers in the dataset have notably low meta-review overall assessment scores and were nevertheless accepted. The number of low-scoring papers accepted by each conference are shown in Table 2.

venue	lowest score	#main	#findings
EMNLP 2024	2	2	9
COLING 2025	1	1	-
	2	12	-
NAACL 2025	2	1	3

Table 2: The lowest meta-review overall assessment scores in each venue (**lowest score**), the number of main conference acceptances (**#main**) and findings acceptances (**#findings**). COLING 2025 does not have findings proceedings.

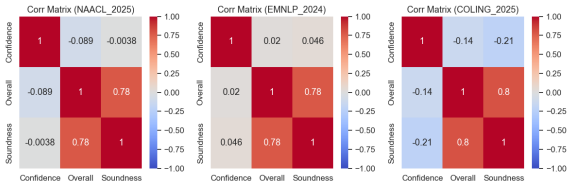


Figure 6: Review-level Pearson correlations between reviewer confidence, overall score, and soundness by venue.

Scoring Behavior We are interested in scoring behavior to explore if evaluations across venues are comparable and to understand whether confidence and soundness relate to overall scores. We want to answer the following research question by exploring scoring behavior. **RQ**: How do soundness and reviewer confidence relate to the overall score, and does this relationship change by venue?

Figure 6 shows the correlation between the various scores. In all three venues, there is a strong positive correlation between overall score and soundness (NAACL (0.78), EMNLP (0.78), COLING (0.80)), confirming that reviewers' overall assessments are closely tied to their evaluation of a paper's correctness. However, the relationship between reviewer confidence and scoring varies more across venues. At NAACL and EMNLP, the correlations between confidence and both overall and soundness scores are near zero (ranging from -0.09 to 0.046), indicating a weak relationship. In contrast, COLING shows a moderate negative correlation between confidence and soundness (-0.21) and overall score (-0.14), suggesting that more confident reviewers may tend to be slightly more critical. This variation at COLING could point to either differences in reviewer behavior or a possible mismatch between self-assessed confidence and actual scoring patterns in that venue.

3.3 Reviewer Focus Analysis

Review aspects are characteristics of a paper that a reviewer makes a judgment on when evaluating the paper (Lu et al., 2025). The distribution of review aspects measures the importance of different paper dimensions to the reviewers, useful for quantifying differences in review behavior across domains and time.

In this section, we analyze reviewer focus using the NLPeer dataset, the ICLR dataset, and the newly collected ARR-Dv2 dataset. We present a cross-temporal and cross-venue cpmparison to show how reviewer focus differs across different settings.

Aspect Tagging We perform an aspect-based analysis of the newly collected data using the aspect scheme and aspect tagger by Lu et al. (2025). Their aspect scheme includes two sets of labels based on granularity: the **COARSE** and **FINE** label sets. Table 8 in Appendix C gives examples of aspects for both.

Table 3 shows the average number of aspects identified per review for each dataset. The number of aspects per review is generally similar for both the **COARSE** and **FINE** label sets.

dataset	COARSE	FINE
NLPeer ARR22	6.28	9.74
NLPeer COLING2020	7.62	13.42
ARR-Dv2 EMNLP 2024	6.35	10.71
ARR-Dv2 COLING 2025	6.31	11.12

Table 3: Average number of aspects identified per review for each dataset, using the **COARSE** and **FINE** label sets.

Figure 7 shows the five most frequent **COARSE** and **FINE** aspects identified in the reviews for each venue. We also include ICLR 2025 data[3] for comparison, which is not part of the ARR-Dv2 data. For **COARSE** aspects, older data (NLPeer ARR-22 and COLING 2020) show almost identical distributions to the new data. This suggests that the reviewer's focus remains largely stable over time. One notable exception is NLPeer COLING 2020, which includes the aspect DDDDEI (Definition, Description, Detail, Discussion, Explanation, Interpretation), suggesting an emphasis on explainability and interpretability. In all venues, the most frequent **COARSE** aspects are Methodology, Result, Data/Task, and Presentation. There is no major difference between ACL venues and ICLR 2025,

[3]Publicly available through the OpenReview API.

indicating similar reviewer focus between the NLP and ML communities at a COARSE level.

In terms of **FINE** aspects, the overall observation is similar to the COARSE aspect, that the reviewer's focus remains stable across time and venues. Common aspects such as Model, Data, Experiment, and Method are consistently mentioned across both earlier and more recent venues. Note that ICLR 2025 has a more balanced aspect distribution. This may suggest that ICLR reviewers place more uniform emphasis across different aspects.

Figure 8 shows the five most frequent **FINE** aspects related to Methodology in the reviews. Reviewers consistently focus on Model and Method, especially in EMNLP 2024 and NAACL 2025. Compared to ACL venues, ICLR 2025 shows a slightly lower emphasis on Model and Method, with a more balanced distribution overall.

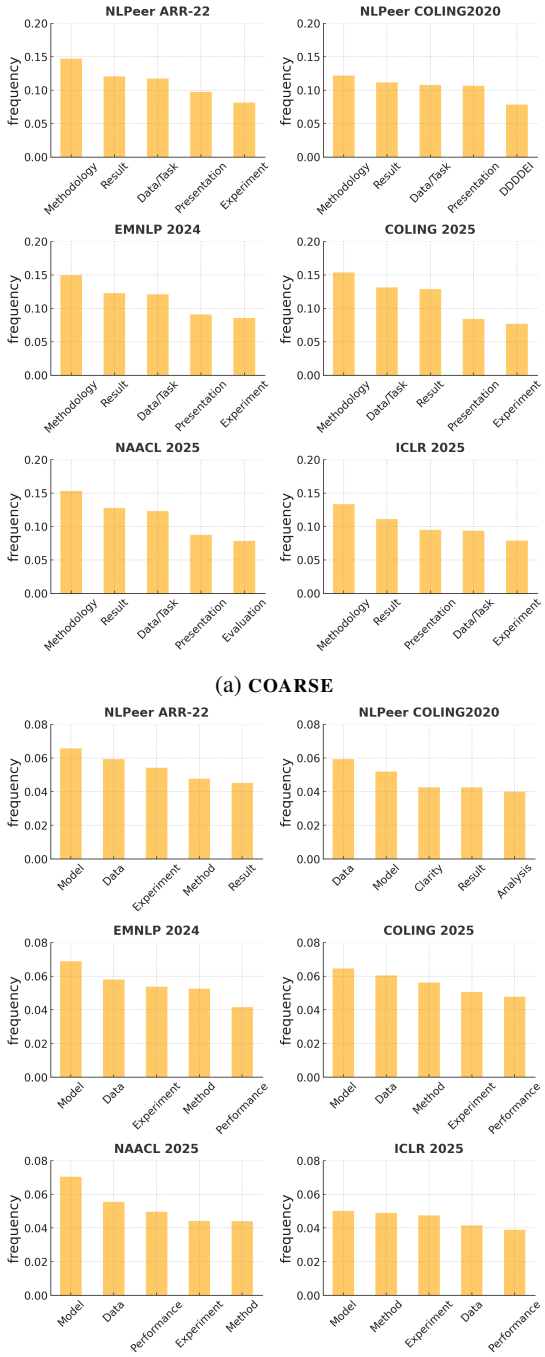
4 Review Score Prediction

In review score prediction, the goal is to predict the numerical review score for a given review report plus potential additional artifacts, such as the review template and scoring rubrics. Solving this task requires modeling the relationship between the review text and the overall reviewer opinion which may vary depending on the reviewing venue and reviewing habits. Here, review score prediction models can serve as assistant tools for reviewers to make more consistent and accurate scoring decisions.

ARR-Dv2 allows us to extend previous work in two important ways: first, the review score prediction also incorporates discussion data, i.e., comments by authors and reviewers regarding the initial review and clarification of those comments. This allows us to explore whether incorporating rebuttal data improves the accuracy of review score prediction, which provides insights into how rebuttals influence reviewers' scoring. Second, we can now perform meta-review score prediction to shine a light on the relationship of scores and text for meta-reviews, too.

4.1 Experimental Settings

For all experiments, we provide the review text and the review rubrics of the venue in the prompt. In addition to predicting scores, we instruct the model to generate justifications for its score predictions to encourage consistency (see Tables 9 and 10 in Appendix D.1). All experiments are in a zero-shot setting, using `Qwen3-14B` (Yang et al., 2025). We chose this model primarily because it is a recent, open-source model with strong performance.



(a) **COARSE**

(b) **FINE**

Figure 7: The 5 most frequent aspects in the reviews for each dataset using the **COARSE** and **FINE** label sets.

Individual Review Score Prediction uses two different setups: the individual review (**review**) and the individual review plus its rebuttal (**review+rebuttal**).

input	review field	acc@0.0	acc@0.5	sum	RMSE	f1
meta-review	Overall Assessment	0.7153	-	-	0.5493	0.3820
review	Soundness	0.3830	0.4697	0.8527	0.5378	0.1768
	Overall Assessment	0.3642	0.4551	0.8192	0.5757	0.1587
review+rebuttal	Soundness	0.3266	0.4182	0.7448	0.6598	0.1140
	Overall Assessment	0.3646	0.3859	0.7505	0.6583	0.1418

Table 4: Results of review score prediction under different settings. **sum** denotes the sum of **acc@0.0** and **acc@0.5**. Note that the meta-review rubrics do not have 0.5-point scoring, so **acc@0.5** and **sum** are marked as "-."

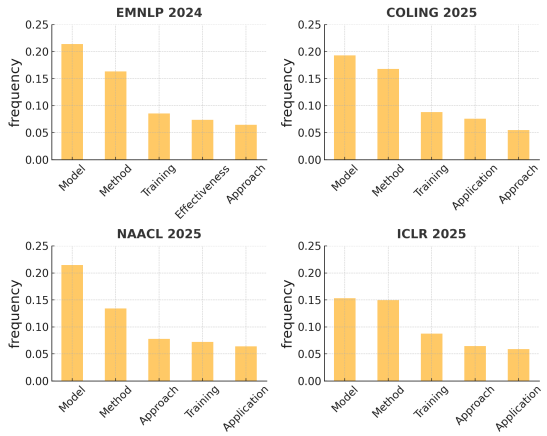


Figure 8: The 5 most frequent **FINE** aspects of Methodology in the reviews.

Meta-Review Score Prediction is the major extension of this work, where we do not have any comparison data. Therefore, we use the same setup as with the individual review score prediction and an initial test.

Evaluation is based on the exact match accuracy (**acc@0.0**) and the percentage of score predictions that differ from the human-assigned score by 0.5 points (**acc@0.5**). We also report **sum**, which denotes the sum of **acc@0.0** and **acc@0.5**. These metrics offer a clear and interpretable view of model performance. **acc@0.0** is the most straight forward and strict metric, while **acc@0.5** allows for a small, reasonable margin of error. Taking them together, **sum** provides a balanced view. We calculate the rooted mean squared error (**RMSE**) and macro F1 score of the predictions. See Appendix D.1 for more details on the experiment settings.

4.2 Results

Individual Review Score Prediction results are shown in Table 4. The model is strong in terms of **sum**, with over 80% of its predictions falling within 0.5 points of the human-assigned scores. The model performs slightly better on Soundness than Overall Assessment score prediction. Our RMSE and f1 scores are slightly lower than those reported by Dycke et al. (2023), but it is worth noting that their results are based on fine-tuned models, whereas ours are in the zero-shot setting. To better interpret the results, see Tables 11, 12 and 13 in Appendix D.2 for results broken down by venues, and random and majority baselines.

Interestingly, incorporating the rebuttal into the input does not seem to improve the model's performance. One possible explanation is that rebuttals tend to be overly positive, which may mislead the model. To better understand this, we conducted a deeper analysis of the review data. We identified two types of reviews: those that include reviewer responses–where the reviewer replies, discusses with the author, or acknowledges the author's responses and those without reviewer responses. We then calculated the model's performance on reviews that include reviewer responses to see if these interactions help the model to better predict review scores. There are 878 out of 1923 reviews that have reviewer responses. Table 5 shows the model's performance on reviews with reviewer responses. Using review+rebuttal as input achieves higher **acc@0.0** on Overall Assessment score prediction than using the individual review alone. This suggests that reviewer responses are valuable for reducing the positivity bias of the rebuttal. Table 14 in the Appendix shows more detailed results.

input	review field	acc@0.0
review	Soundness	0.4002
	Overall Assessment	0.3830
review+rebuttal	Soundness	0.3531
	Overall Assessment	0.4237

Table 5: Results of review score prediction on reviews with reviewer responses.

Meta-Review Score Prediction is also shown in Table 4. The model achieves quite good per-

formance, with over 0.7 accuracy. Please note that Meta-Reviews only have integer-based scores, rather than decimal-based scores. Also, the model was not fine-tuned for the task, which would improve the results.

Positivity Bias results are in Table 6, which show the percentage of cases where the model's predicted score is higher or lower than the human-assigned score. When rebuttal data is included, the model shows a stronger positivity bias than when only the review data is included. It further supports our hypothesis that rebuttals introduce a positivity bias in model predictions.

In contrast, for meta-reviews, the model slightly favors giving lower scores compared to the human-assigned ones. Notably, for COLING 2025, the model shows a strong positivity bias across all tasks and input types (see Table 7). This discrepancy may stem from incomplete data[4]. Another possible explanation is that the review score distribution in COLING 2025 is relatively low compared to other venues (see Figures 4 and 5 in Section 3.2). As a result, the model tends to overestimate the score, leading to a more noticeable positivity bias. See Table 15 in the Appendix for more detailed results.

The analysis above assumes that human-assigned scores are the gold standard. However, there are cases where human-assigned scores do not fully align with their review texts. From this perspective, the observed positivity bias in model predictions may also be interpreted as an indicator of such mismatches between the review text and human-assigned scores. To further investigate this, we conducted a small-scale qualitative analysis on cases with large discrepancies between human-assigned and LLM-predicted scores (see Table 16 in the Appendix). Our findings suggest that, while in some cases the gap indeed reflects a mismatch between the review text and the human-assigned scores, in most cases the model's predictions still tend to be overly positive.

5 Conclusion and Future Work

We presented a data collection of peer review data that considerably extends previous data collections in terms of both data and additional artifacts, including data from the rebuttal phase and meta-reviews.

We also presented an analysis of the data regarding similarities and differences between the various venues, for example, on the link between the overall score and soundness score. Additionally, we analyzed the correlation between confidence, soundness, and overall score. One encouraging result for our community is that confidence is not correlated with either soundness or the overall score.

We also analyzed the data in comparison to previous data collections, such as NLPeer, regarding both coarse- and fine-grained aspects, and found that the data collections are comparable. ARR-Dv2 can also be regarded as an extension of NLPeer. Last, but not least, we conducted initial experiments on review score prediction, observing that the inclusion of rebuttal information aids in score prediction and that score prediction is successful, especially when considering a divergence of 0.5 points.

We also extended the review score prediction to the task of meta-review score prediction, which also shows promising initial results, on which we can build in the future.

Future work includes the collection of unpublished data, as the current data collections show a strong bias towards published papers. Additionally, we are able to tackle various research questions and peer review tasks outlined by Kuznetsov et al. (2024b), such as (meta-) review summarization and providing feedback to reviewers on their scores based on their written reviews.

input	review field	+%	-%
meta-review	Overall Assessment	14.16	14.31
review	Soundness	33.65	28.06
	Overall Assessment	34.17	29.41
review+rebuttal	Soundness	55.47	11.87
	Overall Assessment	47.50	16.04

Table 6: Proportions of outputs where the LLM predicted score is higher (**+%**) or lower (**-%**) than the human-assigned score.

input	review field	+%	-%
meta-review	Overall Assessment	47.06	2.94
review	Soundness	48.44	17.19
	Overall Assessment	59.38	14.06
review+rebuttal	Soundness	75.00	3.12
	Overall Assessment	71.88	14.06

Table 7: Proportions of outputs where the LLM predicted score is higher (**+%**) or lower (**-%**) than the human-assigned score for COLING 2025.

[4]COLING 2025 includes both direct and ARR submissions, and we only have access to ARR submissions.

Limitations

The data analysis and review score prediction experiments may be influenced by potential bias. The current dataset is not comprehensive, as a number of reviews are missing due to constraints on reviewer consent, and our dataset only includes reviews for accepted papers. As a result, the statistics may overrepresent characteristics specific to accepted submissions, and the experimental results may not accurately reflect the models' true predictive capability across a broader range of reviews.

Ethical Considerations

This work is associated with several ethical issues. First, the data collection itself and its publication come with their own ethical and legal challenges. We addressed those in the paper. Second, we are not promoting the automation of the peer reviewing process. On the contrary, science requires feedback from peers, and the *peer* reviewing process is at the core of this feedback mechanism. What we are suggesting is to use technical *support* to reduce the workload on the reviewers, while keeping the human reviewer in the loop.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 67–93. Association for Computational Linguistics.

Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2021. Peerassist: leveraging on paper-review interactions to predict peer review decisions. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23*, pages 421–435. Springer.

Aliaksandr Birukou, Joseph R. Wakeling, Claudio Bartolini, Fabio Casati, Maurizio Marchese, Katsiaryna Mirylenka, Nardine Osman, Azzurra Ragone, Carles Sierra, and Aalam Wassef. 2011. Alternatives to peer review: Novel approaches for research evaluation. *Frontiers in Computational Neuroscience*, volume 5 - 2011.

Lutz Bornmann. 2011. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Saman Ebadi, Hassan Nejadghanbar, Ahmed Rawdhan Salman, and Hassan Khosravi. 2025. Exploring the impact of generative ai on peer review: Insights from journal reviewers. *Journal of Academic Ethics*, pages 1–15.

Gustavo Lúcius Fernandes and Pedro OS Vaz-de Melo. 2024. Enhancing the examination of obstacles in an automated peer review system. *International Journal on Digital Libraries*, 25(2):341–364.

Ahmed M Hanafi, Mohamed M Al-mansi, Omar A Al-Sharif, et al. 2025. Generative ai in academia: A comprehensive review of applications and implications for the research process. *International Journal of Engineering and Applied Sciences-October 6 University*, 2(1):91–110.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Minhao Jiang, Ken Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024.

Does data contamination make a difference? insights from intentionally contaminating pre-training data for language models. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

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.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. 2024a. What can natural language processing do for peer review? *Preprint*, arXiv:2405.06563.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. 2024b. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

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

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel Mcfarland, and James Y. Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29575–29620. PMLR.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292.

Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying aspects in peer reviews. *arXiv preprint arXiv:2504.06910*.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How much are llms contaminated? A comprehensive survey and the llmsanitize library. *CoRR*, abs/2404.00699.

Ana Carolina Ribeiro, Amanda Sizo, Henrique Lopes Cardoso, and Luís Paulo Reis. 2021. Acceptance decision prediction in peer-review through sentiment analysis. In *EPIA Conference on Artificial Intelligence*, pages 766–777. Springer.

D Sculley, Jasper Snoek, and Alex Wiltschko. 2018. Avoiding a tragedy of the commons in the peer review process. *arXiv preprint arXiv:1901.06246*.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*.

Linda Willems. 2022. New dataset offers unique insights into peer review. Accessed: 2025-06-06.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2189–2198.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan and Pengfei Liu. 2022. Kid-review: knowledge-guided scientific review generation with oracle pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11639–11647.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025. ReC^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *CoRR*, abs/2503.08569.

A More on Data Collection

We use the OpenReview Python package to collect the data. The data collection process adheres to the Yes-Yes-Yes protocol (Dycke et al., 2022), which involves two key license agreements: **Peer Reviewer Content License Agreement** (hereinafter referred to as the *reviewer agreement*) and **Blind Submission License Agreement** (hereinafter referred to as the *author agreement*). The main steps are as follows:

- **Obtain reviewer agreements**
  Use `or_client.get_all_notes(invitation=...)`, where `invitation` is the link to the reviewer agreement in OpenReview.
- **Retrieve accepted papers**
  Directly from the ACL Anthology Repository.
- **Collect submission details**
  Use `or_client.get_all_notes(invitation=...)`, where `invitation` is the link to the submissions in OpenReview, including reviews, rebuttals, and the author agreements.
- **Filter the submissions**
  Iterate through all submissions and select those that meet all of the criteria of the Yes-Yes-Yes protocol.

B Data analysis - soundness and overall score by confidence

Figure 9 shows an analysis of our data with respect to confidence and soundness. To maintain readability, the y-axis is skewed in each of these plots. But we can see that we have very few papers with a confidence of 1 or 2. The majority of the reviewers give their confidence as either 3 or 4, and a few reviewers claim a confidence of 5. We can also see that the distribution focuses around a soundness score of 3, indicating that the distribution approximately follows a Gaussian Distribution.

Figure 10 shows a similar distribution for confidence in relation to the overall score. Few reviewers give a low confidence rating of 1 or 2, while the majority note a confidence rating of 3 or 4, with a smaller group giving a confidence rating of 5. Again, the scores resemble a Gaussian distribution, with the majority of scores being 3.

Both Figures 9 and 10 show that contrary to the previous data collection our dataset also contains papers that would probably be rejected for the conference, as overall scores of 1 and 2 (a total of 80 papers for NAACL, 182 papers for EMNLP, and 31 papers for COLING) would probably be a clear reject, while papers with 3 (217 papers for NAACL, 657 papers for EMNLP and 26 papers for COLING) are borderline and papers with an overall score of 5 (3 papers for NAACL, 6 papers for EMNLP) are very likely to be accepted. Across all three NLP venues, both the soundness and overall score distributions per reviewer confidence show a strong central tendency in which the vast majority of reviews, regardless of venue or confidence level, assign moderate scores, with very few ratings for 1 or 5.

C More on Data Analysis

Table 8 shows examples of aspects derived by Lu et al. (2025).

D More on Review Score Prediction

D.1 More on Experiment Settings

For all experiments, we set `temperature=0.5`, `max_tokens=1024`, and `seed=2266`. We used 4-bit quantization for model inference. All experiments were done on either an NVIDIA A100 or H100 GPU.

Tables 9 and 10 show the prompts used in the experiments.

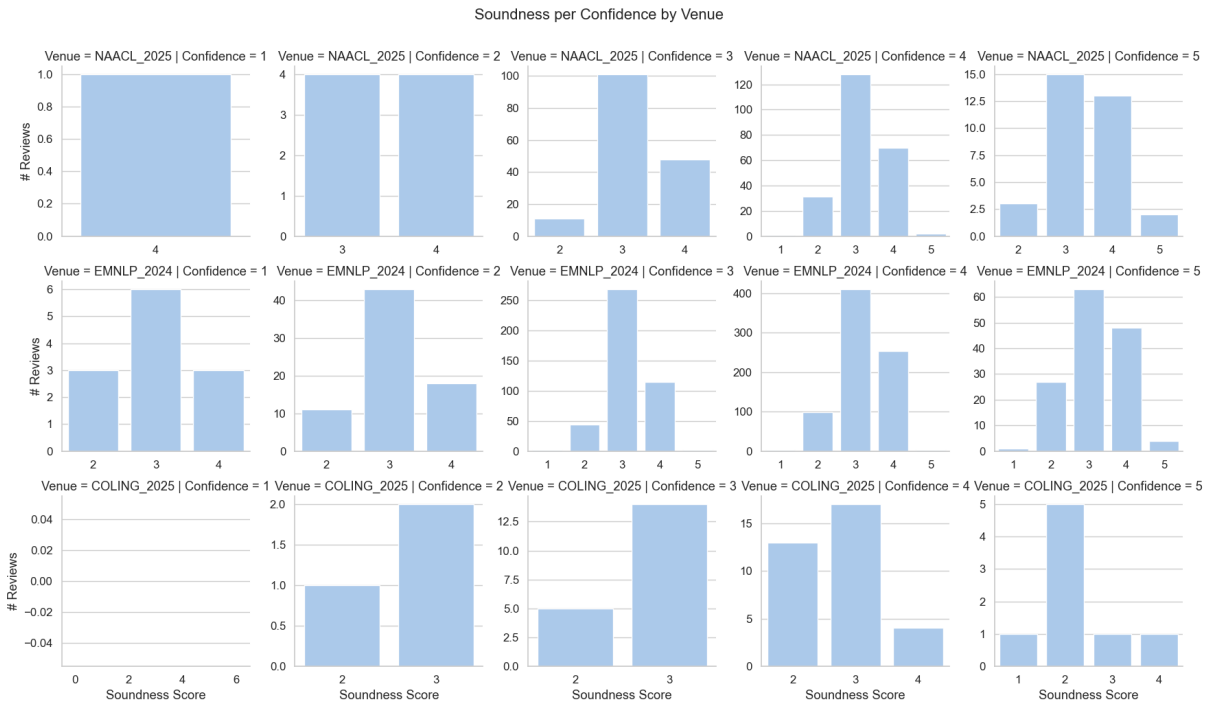


Figure 9: The bars show the average scores for *reviews*, with respect to *confidence* and *soundness* for the papers in our data collection.

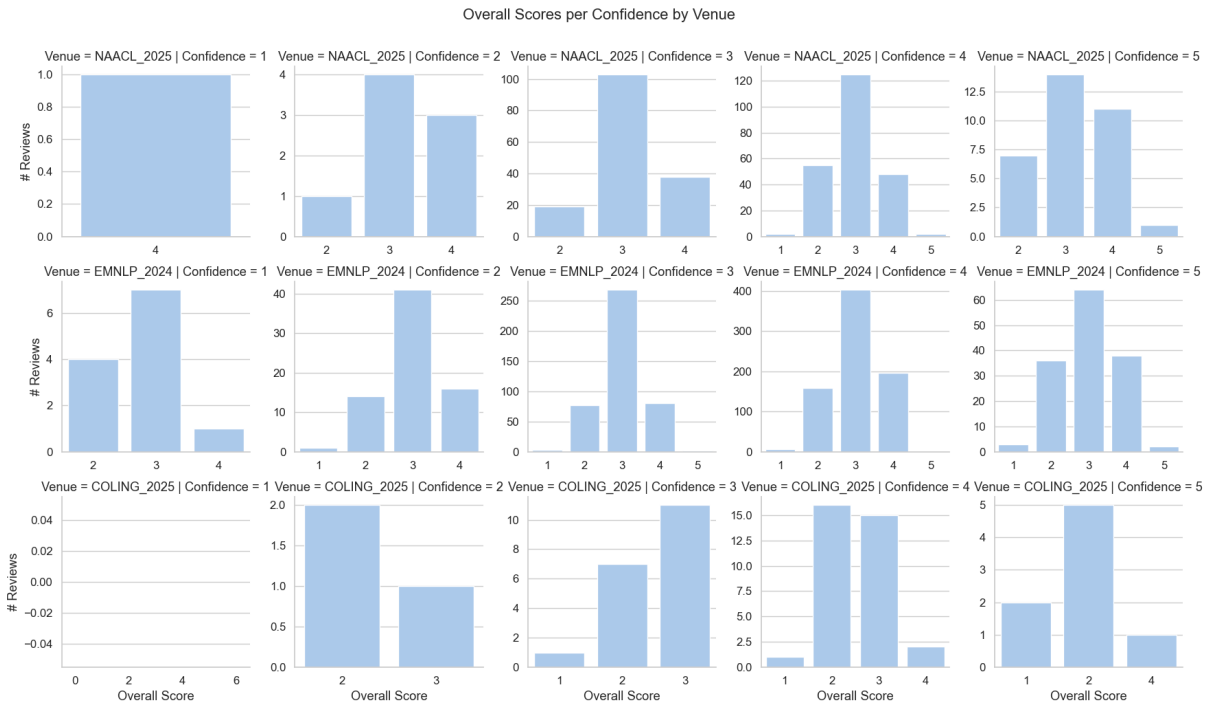


Figure 10: The bars show the average scores for *reviews*, with respect to *confidence* and *overall* score for the papers in our data collection.

D.2 More on Results

Table 11 shows the results of review score prediction broken down by venues. Table 12 shows the random baselines for review score prediction, and Table 13 shows the majority baselines.

Table 14 shows the results of review score prediction on reviews with reviewer responses broken down by venues.

COARSE	FINE
Contribution	Contribution
DDDDEI	Definition, Description, Detail, Discussion, Explanation, Interpretation
IJMV	Intuition, Justification, Motivation, Validation
Novelty	Innovation/Novelty/Originality
Presentation	Clarity, Figure, Grammar, Presentation, Typo
Related Work	Citation/Literature/Related Work
Significance	Impact, Importance, Significance
Ablation	Ablation
Analysis	Analysis
Comparison	Comparison
Data/Task	Annotation, Benchmark, Data, Task
Evaluation	Evaluation, Metric
Experiment	Experiment
Methodology	Algorithm, Implementation, Method
Theory	Theory
Result	Findings, Improvement, Performance, Result

Table 8: Examples of aspects in the COARSE and FINE label sets (Lu et al., 2025).

Given a paper metareview and the corresponding metareview form, recommend appropriate Overall Assessment score that the metareviewer might consider giving based on their review.

Instructions:
1. Read the review text carefully.
2. Analyze the review and recommend an Overall Assessment score.
3. Provide a brief explanation for the recommended score.
4. Output a json dictionary: `{"overall_assessment_score": ..., "overall_assessment_justification": ...}`
5. Output only the json dictionary and follow the json schema exactly, no extra keys in the output.

Review form:

Overall Assessment
5 = The paper is largely complete and there are no clear points of revision
4 = There are minor points that may be revised
3 = There are major points that may be revised
2 = The paper would need significant revisions to reach a publishable state
1 = Even after revisions, the paper is not likely to be publishable at an *ACL venue

Table 9: The prompt used for the meta-review score prediction task.

Table 15 shows the positivity results broken down by venues.

Table 16 shows examples of human-assigned and LLM-predicted review scores with large discrepancies.

Given a paper review and the corresponding review form, recommend appropriate Soundness and Overall Assessment scores that the reviewer might consider giving based on their review.

Instructions:
1. Read the review text carefully.
2. Analyze the review and recommend a Soundness score and an Overall Assessment score.
3. Provide a brief explanation for each recommended score.
4. Output a json dictionary: `{"soundness_score": ..., "soundness_justification": ..., "overall_assessment_score": ..., "overall_assessment_justification": ...}`
5. Output only the json dictionary and follow the json schema exactly, no extra keys in the output.

Review form:

Soundness
How sound and thorough is this study? Does the paper clearly state scientific claims and provide adequate support for them? For experimental papers: consider the depth and/or breadth of the research questions investigated, technical soundness of experiments, methodological validity of evaluation. For position papers, surveys: consider the current state of the field is adequately represented, and main counter-arguments acknowledged. For resource papers: consider the data collection methodology, resulting data & the difference from existing resources are described in sufficient detail. Please adjust your baseline to account for the length of the paper.

5 = Excellent: This study is one of the most thorough I have seen, given its type.
4.5
4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
3.5
3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
2.5
2 = Poor: Some of the main claims/arguments are not sufficiently supported. There are major technical/methodological problems.
1.5
1 = Major Issues: This study is not yet sufficiently thorough to warrant publication or is not relevant to ACL.

Overall Assessment
Would you personally like to see this paper presented at an *ACL event that invites submissions on this topic? For example, you may feel that a paper should be presented if its contributions would be useful to its target audience, deepen the understanding of a given topic, or help establish cross-disciplinary connections. Note: Even high-scoring papers can be in need of minor changes (e.g. typos, non-core missing refs, etc.).

5 = Top-Notch: This is one of the best papers I read recently, of great interest for the (broad or narrow) sub-communities that might build on it
4.5
4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it
3.5
3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions
2.5
2 = Revisions Needed: This paper has some merit, but also significant flaws, and needs work before it would be of interest to the community
1.5
1 = Major Revisions Needed: This paper has significant flaws, and needs substantial work before it would be of interest to the community
0 = This paper is not relevant to the *ACL community (for example, is in no way related to natural language processing)

Table 10: The prompt used for the individual review score prediction task.

input	venue	review field	acc@0.0	acc@0.5	sum	RMSE	f1
meta-review	EMNLP 2024	Overall Assessment	0.7354	-	-	0.5319	0.4971
	COLING 2025	Overall Assessment	0.5000	-	-	0.7071	0.3543
	NAACL 2025	Overall Assessment	0.7031	-	-	0.5589	0.3845
review	EMNLP 2024	Soundness	0.3832	0.4658	0.8490	0.5417	0.1785
		Overall Assessment	0.3592	0.4552	0.8144	0.5803	0.1563
	COLING 2025	Soundness	0.3438	0.4531	0.7969	0.5962	0.2297
		Overall Assessment	0.2656	0.4531	0.7188	0.7126	0.1478
	NAACL 2025	Soundness	0.3880	0.4850	0.8730	0.5154	0.1825
		Overall Assessment	0.3949	0.4550	0.8499	0.5357	0.1849
review+rebuttal	EMNLP 2024	Soundness	0.3298	0.4232	0.7530	0.6490	0.1230
		Overall Assessment	0.3691	0.3867	0.7558	0.6506	0.1451
	COLING 2025	Soundness	0.2188	0.4219	0.6406	0.8173	0.1622
		Overall Assessment	0.1406	0.4219	0.5625	0.8429	0.0664
	NAACL 2025	Soundness	0.3318	0.4014	0.7332	0.6667	0.1368
		Overall Assessment	0.3828	0.3782	0.7610	0.6529	0.1578

Table 11: Results of review score prediction under different settings. **sum** denotes the sum of **acc@0.0** and **acc@0.5**. Note that the meta-review rubrics do not have 0.5-point scoring, so **acc@0.5** and **sum** are marked as "-."

input	venue	review field	acc@0.0	acc@0.5	acc@1.0
meta-review	EMNLP 2024	Overall Assessment	0.2042	-	0.3625
	NAACL 2025	Overall Assessment	0.2396	-	0.3698
review	EMNLP 2024	Soundness	0.0996	0.2007	0.2091
		Overall Assessment	0.1186	0.1811	0.1993
	NAACL 2025	Soundness	0.1129	0.1820	0.1935
		Overall Assessment	0.1083	0.1613	0.2166

Table 12: Random baselines for review score prediction. Note that the meta-review rubrics do not have 0.5-point scoring, so **acc@0.5** is marked as "-."

input	venue	review field	acc@0.0	acc@0.5	acc@1.0
meta-review	EMNLP 2024	Overall Assessment	0.6354	-	0.3417
	NAACL 2025	Overall Assessment	0.6615	-	0.3177
review	EMNLP 2024	Soundness	0.3060	0.3446	0.2933
		Overall Assessment	0.3179	0.3811	0.2596
	NAACL 2025	Soundness	0.3410	0.2995	0.3111
		Overall Assessment	0.3203	0.3871	0.2627

Table 13: Majority baselines for review score prediction. Note that the meta-review rubrics do not have 0.5-point scoring, so **acc@0.5** is marked as "-."

venue	input	review field	acc@0.0
EMNLP24	review	Soundness	0.4063
		Overall Assessment	0.3852
	review+rebuttal	Soundness	0.3638
		Overall Assessment	0.4147
COLING25	review	Soundness	0.4615
		Overall Assessment	0.3462
	review+rebuttal	Soundness	0.3077
		Overall Assessment	0.1923
NAACL25	review	Soundness	0.3696
		Overall Assessment	0.3804
	review+rebuttal	Soundness	0.3207
		Overall Assessment	0.4891

Table 14: Results of review score prediction on reviews with reviewer responses.

input	venue	review field	+%	-%
meta-review	EMNLP24	Overall Assessment	12.08	14.37
	COLING25		47.06	2.94
	NAACL25		13.54	16.15
review	EMNLP24	Soundness	32.32	29.36
		Overall Assessment	33.10	30.98
	COLING25	Soundness	48.44	17.19
		Overall Assessment	59.38	14.06
	NAACL25	Soundness	35.80	25.40
		Overall Assessment	33.95	26.56
review+rebuttal	EMNLP24	Soundness	54.46	12.56
		Overall Assessment	45.82	17.26
	COLING25	Soundness	75.00	3.12
		Overall Assessment	71.88	14.06
	NAACL25	Soundness	55.92	10.90
		Overall Assessment	49.42	12.30

Table 15: Proportions of outputs where the LLM predicted score is higher (**+%**) or lower (**-%**) than the human-assigned score.

Human review	Human		LLM		LLM justifications
	S	OA	S	OA	
Strengths: The paper addresses formatting issues of web mathematics datasets and proposes a reasonable pipeline. *Weaknesses:* Missing details on key components and no validation across domains.	2	1	3	3	**Soundness:** Reasonable approach but lacks detail and cross-domain validation. **Overall:** Addresses a specific problem but limited generalizability; minor revisions could improve impact.
Strengths: Well-organized and smoothly written manuscript; dataset focuses on complex scenarios. *Weaknesses:* None.	2.5	3	5	5	**Soundness:** Thorough and sound research with clear writing. **Overall:** High-quality, valuable dataset of broad interest—top-notch contribution.
Strengths: K2Q increases diversity and complexity of VQA questions, improving robustness. Covers multiple domains and allows flexible expansion. *Weaknesses:* Some unnatural questions; domain coverage limitations.	3	2	4	4	**Soundness:** Technically sound dataset generation method with clear claims; some refinement needed. **Overall:** Significant contribution to dataset creation; limitations do not detract from value.
Strengths: Multidimensional dataset for summarization evaluation; detailed annotation; systematic LLM comparison. *Weaknesses:* Missing automatic metrics and slightly imbalanced data sources.	2	1.5	3.5	3.5	**Soundness:** Comprehensive description and evaluation, but missing key metrics. **Overall:** Useful dataset with room for improvement in size and balance.
Strengths: Clearly written; effective experiments. *Weaknesses:* Limited novelty; assumptions not validated; mechanism unclear.	2	1	3	3	**Soundness:** Clear problem and solid experiments, but lacks novelty and validation depth. **Overall:** Reasonable contribution needing minor revisions to strengthen claims.
Strengths: Systematic study on temperature parameter's effect. *Weaknesses:* Obvious findings; ignores other parameters (e.g., top-p).	1	1	3	3	**Soundness:** Supports main claims but lacks depth on related parameters. **Overall:** Reasonable contribution, though limited in novelty and breadth.
Strengths: Investigates stopword and lemmatization effects. *Weaknesses:* Lacks dataset details and regex explanation.	1	1	3	3	**Soundness:** Clear claim but insufficient methodological detail. **Overall:** Modest contribution; needs clarification for better reproducibility.
Strengths: Includes ablation and clear visualizations. *Weaknesses:* Lacks comparison with post-2022 work and more challenging datasets.	2.5	1	3.5	3	**Soundness:** Adequate support for claims, but limited scope and depth. **Overall:** Reasonable contribution, but relevance could be improved with newer benchmarks.

Table 16: Examples of human-assigned and LLM-predicted review scores with large discrepancies. S is for Soundness score, and OA is for Overall Assessment score.