

---

# Re<sup>2</sup>: A Consistency-ensured Dataset for Full-stage Peer Review and Multi-turn Rebuttal Discussions

---

**Daoze Zhang\***  
Zhejiang University  
zhangdz@zju.edu.cn

**Zhijian Bao\***  
Zhejiang University  
baozhijian@zju.edu.cn

**Sihang Du**  
Zhejiang University  
dusihang@zju.edu.cn

**Zhiyi Zhao**  
Zhejiang University  
zhaozhiyi@zju.edu.cn

**Kuangling Zhang**  
Zhejiang University  
zhangkl@zju.edu.cn

**Dezheng Bao**  
Zhejiang University  
baodezheng@zju.edu.cn

**Yang Yang†**  
Zhejiang University  
yangya@zju.edu.cn

## Abstract

Peer review is a critical component of scientific progress in the fields like AI, but the rapid increase in submission volume has strained the reviewing system, which inevitably leads to reviewer shortages and declines review quality. Besides the growing research popularity, another key factor in this overload is the repeated resubmission of substandard manuscripts, largely due to the lack of effective tools for authors to self-evaluate their work before submission. Large Language Models (LLMs) show great promise in assisting both authors and reviewers, and their performance is fundamentally limited by the quality of the peer review data. However, existing peer review datasets face three major limitations: (1) limited data diversity, (2) inconsistent and low-quality data due to the use of revised rather than initial submissions, and (3) insufficient support for tasks involving rebuttal and reviewer-author interactions. To address these challenges, we introduce the largest consistency-ensured peer review and rebuttal dataset named Re<sup>2</sup>, which comprises 19,926 initial submissions, 70,668 review comments, and 53,818 rebuttals from 24 conferences and 21 workshops on OpenReview. Moreover, the rebuttal and discussion stage is framed as a multi-turn conversation paradigm to support both traditional static review tasks and dynamic interactive LLM assistants, providing more practical guidance for authors to refine their manuscripts and helping alleviate the growing review burden. Our data and code are available in this repository.

## 1 Introduction

Peer review is a cornerstone in the advancement of scientific research, ensuring that high-value works which are sufficiently novel, credible, and rigorously evaluated can be published. With the rapid surge in submission volume in some fields like Computer Science (CS) or Artificial Intelligence (AI), peer researchers have to bear increasing pressure to review, leading to a shortage of reviewers and an inevitable decline in review quality. In addition to the growing popularity of these disciplines, another major contributor to the submission overload is the repeated resubmission of manuscripts

---

\* Equal contribution.

† Corresponding author.

that fall short of quality standards. This trend is often driven by the lack of effective tools for authors to objectively assess and improve their manuscripts prior to submission, resulting in multiple rounds of submission for the same study. For the above two issues in the research and review stages, Large Language Models (LLMs) hold significant potential to become a powerful assistant to alleviate these problems. Specifically, LLM-based review tools can be used by authors as a pre-submission self-evaluation mechanism, allowing authors to identify and improve the weaknesses in the manuscripts, and thereby enhancing the work quality and reducing resubmission rates. Additionally, LLMs can be directly integrated into the peer review process to support reviewers in generating more specific and constructive feedback — as demonstrated by the ICLR 2025 conference, which has introduced an LLM-powered reviewer assisting system [ICLR, 2025].

For the training of language models with peer review capabilities, the most crucial knowledge foundation that determines the upper bound of the model performance is the real-world high-quality paper review data. Numerous researchers have constructed diverse peer review datasets to support the training and evaluation of LLM-based reviewing assistants. These existing datasets target a range of tasks, including: (1) prediction-oriented tasks such as acceptance prediction and score prediction [Kang et al., 2018, Bharti et al., 2021, Dycke et al., 2022]; (2) generation tasks, including review and meta-review generation [Shen et al., 2021, Zhang et al., 2022, Yuan et al., 2022, Wu et al., 2022, D’Arcy et al., 2023, Jin et al., 2024, Gao et al., 2024, Zhou et al., 2024, Weng et al., 2024, Zhu et al., 2025]; and (3) analytical tasks related to review content, like review action analysis [Kennard et al., 2021, Purkayastha et al., 2023, Bharti et al., 2024]. Despite the above advancements in the field of automated paper review, existing datasets still exhibit the following critical limitations.

From the perspective of data diversity, **the paper sources of these existing datasets are limited to a few conferences, and the amount of data is also relatively small.** Due to the huge differences in the availability and formatting of the review information across different conferences, the majority of existing datasets [Kennard et al., 2021, Bharti et al., 2021, Wu et al., 2022, Zhang et al., 2022, Zhou et al., 2024, Weng et al., 2024, Jin et al., 2024, Zhu et al., 2025] are only based on the review data from the ICLR conference, which is known for its high level of transparency in the review process. A few other works [Kang et al., 2018, Shen et al., 2021, Yuan et al., 2022, Dycke et al., 2022, D’Arcy et al., 2023, Gao et al., 2024] incorporate reviews from additional conferences like NeurIPS and ACL. However, none of them comprehensively capture all publicly available peer reviews from conferences hosted on OpenReview, leading to obvious limitations in both the data source diversity and data scale.

From the viewpoint of data quality, most existing datasets have a fatal problem — **the provided paper content may correspond to the revised version rather than the initial submission.** Obviously, for training and evaluating language models on the task of automated paper review, the paper content fed into models is supposed to be the initial submission that is not revised by the author, as this is the version to which the reviewers’ comments are actually addressed. However, as shown in Tab. 1, most existing datasets (not highlighted in yellow) fail to guarantee that the paper contents are indeed the initial submissions, and contain versions that have been revised by the authors in response to review comments. This discrepancy introduces substantial risks to the coherence and consistency of the review data, undermining the rationality of model training and evaluation for the review-related tasks.

From the task perspective, many existing works mainly **stay on the traditional review generation task, while overlooking the valuable data contained in the rebuttal and discussion stages.** In order to better assist authors in self-evaluation and improving their submissions before formal review, language models are expected not only to generate static review comments, but also to understand the author’s rebuttals and provide further responses. This rebuttal–discussion process constitutes a typical multi-turn conversation task, which holds great potential and research value. However, as shown in Tab. 1, most existing datasets do not contain any information related to rebuttals. For the few datasets [Kennard et al., 2021, Wu et al., 2022, Jin et al., 2024, Tan et al., 2024] involving rebuttals, they often handle the rebuttal and discussion data in a coarse and insufficient manner and also fail to guarantee the consistency of the data, falling far short of supporting research that treats rebuttal and discussion as a well-defined multi-turn conversation task.

To tackle the limitations above, we propose a real-world dataset named  $\text{Re}^2$  for comprehensive academic peer review, to support the training and evaluation of both **review** and **rebuttal** abilities of language models. Our  $\text{Re}^2$  dataset consists of two main parts: (1) the  $\text{Re}^2$ -Review dataset contains 19,926 initial paper submissions and 70,668 review comments from human reviewers collected from OpenReview, covering 24 conferences and 21 workshops from 2017 to 2025; and (2) the  $\text{Re}^2$ -Rebuttal

dataset contains 14,830 initial submissions paired with 53,818 rebuttal and discussions, which are formatted as structured multi-turn conversation to facilitate the training of various language models. By overcoming the huge heterogeneity in data storage formats across different conferences and years, we unify all data into a consistent format to facilitate broad accessibility for peer researchers. Moreover, our  $\text{Re}^2$  dataset covers as many review stages on OpenReview as possible, including initial submissions, reviewer comments, ratings and confidence scores, aspect-specific ratings (e.g., soundness, presentation, contribution), rebuttal–discussion conversations, score changes before and after rebuttal, meta-reviews, and final decisions.

To our knowledge, our  $\text{Re}^2$  is the largest real-world peer review dataset to date, with the broadest inclusion of conferences and the most comprehensive coverage of review stages (details in Tab. 1). In addition, it is the first consistency-ensured dataset to support rebuttal tasks in a multi-turn conversation paradigm. The  $\text{Re}^2$  dataset not only allows traditional static tasks such as computational analysis, acceptance or score prediction, review or meta-review generation, but also enables the training of interactive, chat-based models for further rebuttal and discussion, offering support for authors to self-evaluate and improve their work before submission, which is also helpful in reducing the review burden in the AI community. In summary, our key contributions are as follows:

- To our knowledge, we present the largest real-world consistency-ensured dataset named  $\text{Re}^2$  for peer review and rebuttal-discussion, which features the widest range of conferences and the most complete review stages, including initial submissions, reviews, (aspect) ratings and confidence, rebuttals, discussions, score changes, meta-reviews, and decisions.
- Moving beyond the traditional static review paradigm, we treat the rebuttal and discussion data as a multi-turn conversation task between reviewers and authors, which enables the training and evaluation of dynamic, interactive LLM-based reviewing assistants, offering more practical guidance for authors to self-improve their work before submission.
- We conduct a statistical analysis of the proposed dataset, and experimentally demonstrate its effectiveness in improving the capabilities of language models in peer review and rebuttal scenarios through four review-related tasks.

## 2 The $\text{Re}^2$ Dataset

### 2.1 Data Collection and Processing

**$\text{Re}^2$ -Review Dataset.** The first subset of our proposed dataset is named  $\text{Re}^2$ -Review, which is mainly designed for acceptance prediction, score prediction, and review or meta-review generation tasks. For the  $\text{Re}^2$ -Review dataset, we first utilize the official API of OpenReview to automatically crawl all publicly accessible papers and their full peer review records (including metadata, reviews, rebuttals, etc.) from OpenReview, covering 68 conferences from 2013 to 2025. Afterwards, given that all the review-related tasks must be grounded in the initial submitted manuscripts rather than revised versions, we need to ensure that the paper contents in our dataset are the initial submissions. To achieve this, we comprehensively collect the paper submission deadlines for each conference of different years. Based on the information of deadlines, we then employ web scraping techniques to extract the latest version before the submission deadline from the “Revision History” page of each paper. The final paper set spans 24 conferences and 21 workshops from 2017 onward.

For the extraction of review contents, due to the diversity of conferences and years, these review data come in a wide range of formats. Therefore, the processing logic in existing works, which typically target common conferences such as ICLR or NeurIPS, cannot be directly applied to these review data from all the 45 venues. To address this challenge, we manually audit all data format variations involved, and implement customized extraction logic for each conference-year pair, to achieve automatic and efficient extraction of full-stage review contents across these conferences.

For the paper content, we convert paper formats from PDF into plain text to facilitate the use by the research community. To achieve this, we employ a commercial tool named Doc2X<sup>1</sup>, which outperforms open-source alternatives in terms of recognition accuracy and quality, especially for mathematical formulas. Using this tool, we convert the paper content in our dataset into both LaTeX and Markdown formats for broader accessibility and downstream applications.

<sup>1</sup><https://github.com/NoEdgeAI/doc2x-doc>

Table 1: **Comparison between our Re<sup>2</sup> dataset and existing peer review datasets.** Note that only the datasets in yellow rows can guarantee all provided papers are **initial submissions**, which is critical for the consistency and quality of the review data. For the “Task” column, AP is the abbreviation for Acceptance Prediction, SP for Score Prediction, RA refers to Review Analysis tasks, RG stands for Review Generation, and MG denotes Meta-review Generation. In the three columns representing numbers, “-” means please refer to the original dataset for the number.

Dataset Name	Data Source (Conf.&Year)	Task	Reviews	Aspect Scores	Rebuttal	Score Changes	Meta-reviews	Final Decision	# Paper	# Review Comments	# Rebuttal
PeerRead [Kang et al., 2018]	ICLR 17, ACL 17, NeurIPS 13-17	AP, SP	✓	✓			✓	✓	14,700	10,700	0
DISAPERE [Kennard et al., 2021]	ICLR 19-20	RA	✓		✓				0	506	506
PEERAssist [Bharti et al., 2021]	ICLR 17-20	AP	✓					✓	4,467	13,401	0
MReD [Shen et al., 2021]	ICLR 18-21	MG	✓					✓	7,894	30,764	0
ASAP-Review [Yuan et al., 2022]	ICLR 17-20, NeurIPS 16-19	RG	✓				✓		8,877	28,119	0
NLPeer [Dycke et al., 2022]	ACL 17, ARR 22, COLING 20, CONLL 16	RA	✓	✓					5,672	11,515	0
PRRCA [Wu et al., 2022]	ICLR 17-21	MG	✓		✓	✓	✓	✓	7,627	25,316	-
[Zhang et al., 2022]	ICLR 17-22	RA	✓	✓				✓	10,289	36,453	68,721
ARIES [D’Arcy et al., 2023]	OpenReview	RG	✓						1,720	4,088	0
AgentReview [Jin et al., 2024]	ICLR 20-23	RG, MG	✓	✓	✓	✓	✓	✓	500	10,460	-
Reviewer2 [Gao et al., 2024]	ICLR 17-23, NeurIPS 16-22 (PeerRead & NLPeer)	RG	✓				✓	✓	27,805	99,727	0
RR-MCQ [Zhou et al., 2024]	ICLR 17 (from PeerRead)	RG	✓					✓	14	55	0
ReviewMT [Tan et al., 2024]	ICLR 17-24	RG	✓		✓		✓	✓	26,841	92,017	0
Review-5K [Weng et al., 2024]	ICLR 24	RG	✓					✓	4,991	16,000	0
DeepReview-13K [Zhu et al., 2025]	ICLR 24-25	RG	✓	✓				✓	13,378	13,378	0
<b>Our Re<sup>2</sup></b>	45 venues from 2017 to 2025	RG, MG, AP, SP	✓	✓	✓	✓	✓	✓	<b>19,926</b>	<b>70,668</b>	<b>53,818</b>

**Re<sup>2</sup>-Rebuttal Dataset.** Another subset named Re<sup>2</sup>-Rebuttal is designed as a structured multi-turn conversation dataset based on the rebuttal and discussion data between authors and reviewers. For the construction of our Re<sup>2</sup>-Rebuttal subset, we further filter the review dataset mentioned above to retain only those papers for which the reviewer–author rebuttal stage is publicly accessible, and then process the corresponding rebuttal interactions. Specifically, we organize the authors’ rebuttals, along with the subsequent reviewer–author discussion, into structured multi-turn dialogues. The main challenges encountered here are mainly the following two parts.

First, due to the character limits imposed on each individual response on OpenReview, authors often post multiple consecutive responses during the rebuttal stage, with each response containing only a portion of their overall response (see the blue box in Fig. 1(a) for an example). To convert the rebuttal–discussion process into a well-structured multi-turn dialogue for the training of language models, we concatenate multiple consecutive responses from the same role (e.g., author or reviewer) into a single turn. Then, DeepSeek-R1 [Guo et al., 2025] is further employed to merge the title of

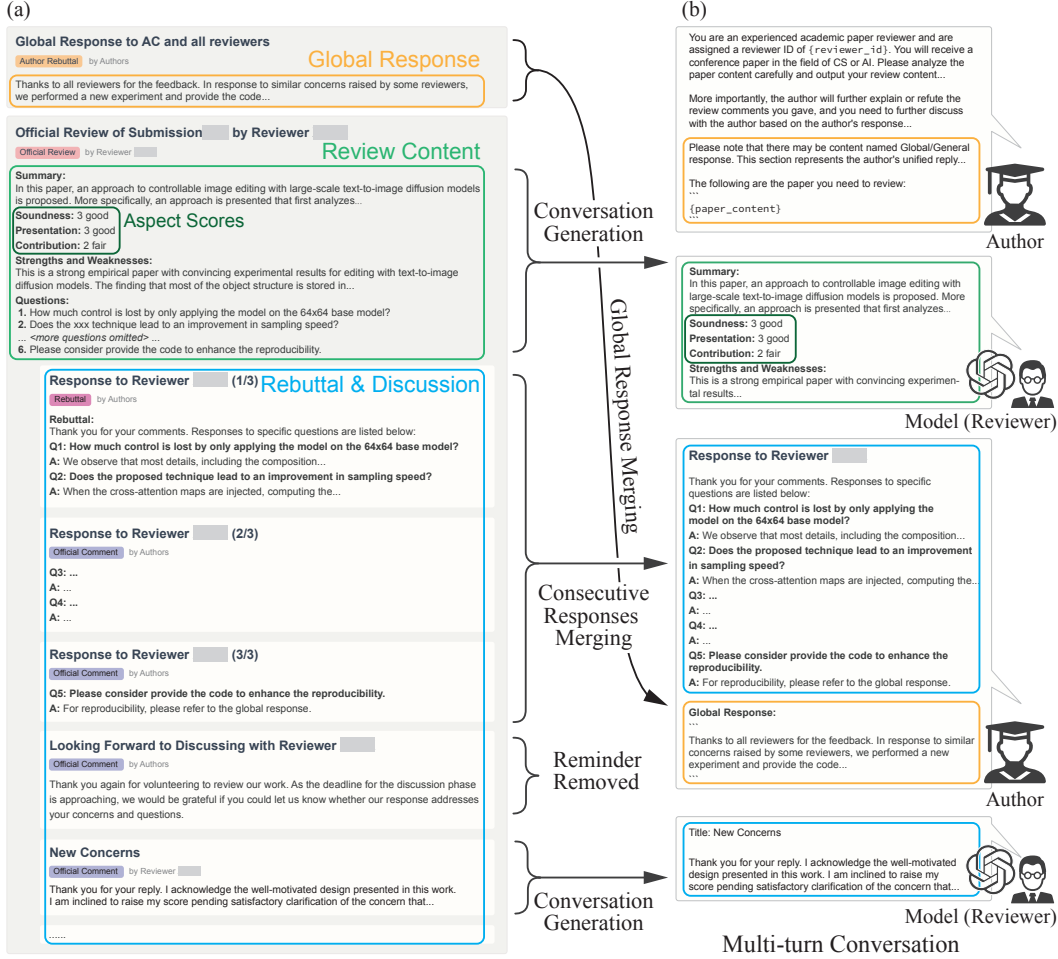


Figure 1: **The conversion from raw review and rebuttal data to multi-turn conversations.** For the raw review data crawled from OpenReview (as shown in sub-figure(a)), we concatenate multiple consecutive responses from the same role (author or reviewer) into a single turn. In cases where the author’s final response is merely a reminder or urging, we adopt a hybrid strategy combining manual inspection and automated methods to identify and remove such reminder responses. As for the global responses, we insert them into the dialogue at the appropriate position, treating it as supplementary reference rather than direct conversation content. Finally, as shown in sub-figure(b), we construct a self-consistent, high-quality, and information-complete multi-turn conversation dataset.

each response, producing a coherent full response along with a unified title. It is worth noting that, in some cases, the final response in a series of consecutive posts from the author is simply a reminder or urging to an unresponsive reviewer. Clearly, such responses should not be concatenated with the previous rebuttal content. To handle this problem, we adopt a hybrid strategy combining manual inspection and automated methods to identify and exclude these follow-up reminders, ensuring the generated multi-turn dialogue data is consistent and high-quality.

Second, when reviewing a submission, sometimes several reviewers may raise similar concerns, and the author will give a unified response to these shared questions through a global or general response (see the orange box in Fig. 1(a) for an example). To construct a complete multi-turn rebuttal–discussion conversation, we incorporate the global responses into the author–reviewer interactions. Specifically, we insert the content of global responses into the dialogue at the appropriate position corresponding to each reviewer’s related comment. However, since global responses may also address concerns raised by other reviewers not involved in the current dialogue thread, we need to take special care to maintain logical consistency. Therefore, we insert global responses using a special format (as illustrated in the orange box in Fig. 1(b)), treating them as referential context

instead of direct replies from the author within the turn. This supplements the dialogue with reference information while preserving the consistency and logical flow of the multi-turn conversation.

## 2.2 Dataset Statistics

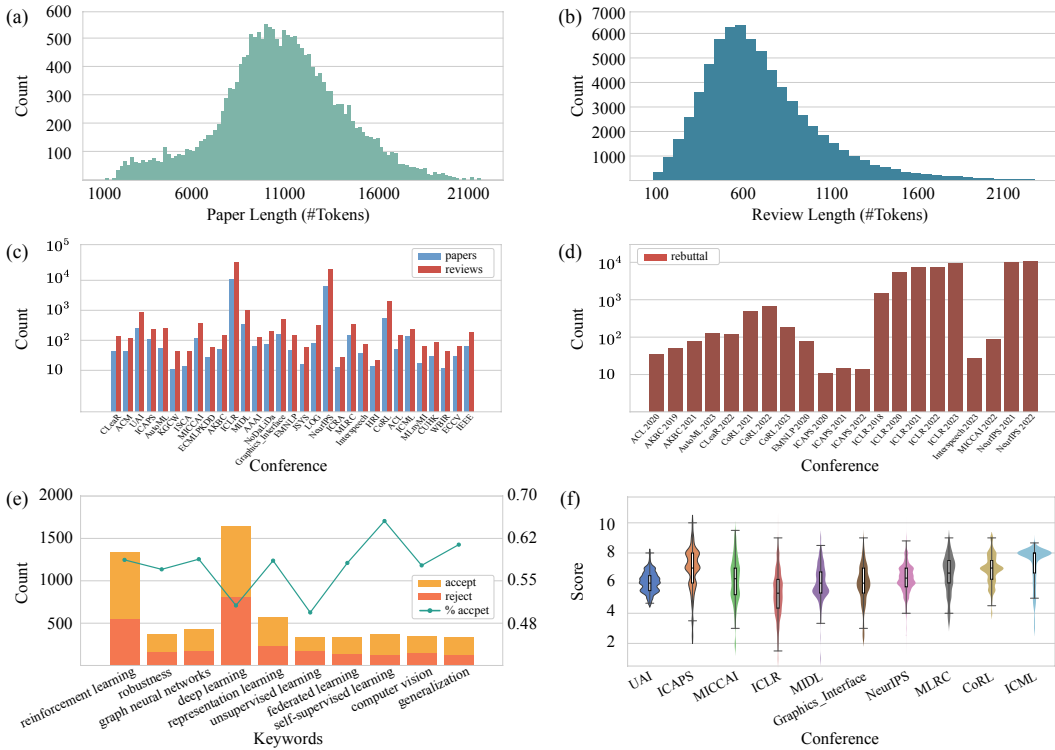


Figure 2: **Statistics of our  $Re^2$  dataset.** (a) Distribution of the length of papers in tokens. (b) Distribution of the length of reviews in tokens. (c) Distribution of the number of papers and reviews in each conference. (d) Distribution of the number of rebuttals in each conference of each year. (e) Submission counts and acceptance proportion across the 10 most frequent keywords. (f) Violin plot (with a box plot inside) of review scores across the top 10 conferences with the most papers.

**Paper, Review, and Rebuttal Distributions.** The distributions of paper and review lengths (in number of tokens) are shown in the histograms in Fig. 2(a) and Fig. 2(b), respectively. Most papers range in length from 6,000 to 16,000 tokens, while the majority of reviews are distributed between 300 and 1,100 tokens. As shown in Fig. 2(c), we count the total number of papers and reviews of each conference. Among them, ICLR and NeurIPS hold the highest proportion of the number of papers and reviews. The distribution of the number of rebuttals for each conference in each year is given in Fig. 2(d), where ICLR and NeurIPS again account for the highest proportion.

**Acceptance and Scores.** As shown in Fig. 2(e), to examine the popularity of different research areas within the AI community, we present the number of papers and the acceptance proportions across the top 10 frequent research keywords. It can be illustrated that the acceptance proportions of “self-supervised learning” and “generalization” rank first and second. Meanwhile, the number of papers about “deep learning” and “reinforcement learning” is the largest, demonstrating that these two are popular research areas in the field of AI. Also, in order to present the distributions of the review scores among different conferences, we utilize the violin plot with a box plot inside, to conduct a statistical analysis of the normalized scores of each conference in Fig. 2(f). Due to the excessive number of conferences, we only show the top 10 conferences with the most papers in detail. In the violins of Fig. 2(f), the bold horizontal line inside the box represents the median score, and the contour width indicates the number of papers of this score. It is worth noting that since Kernel Density Estimation (KDE) is used to smooth the data distribution, it may cause the tail extension of the violin to exceed the range boundary. It can be demonstrated that the median scores of each conference are all between 5 and 7. Additionally, by counting the peaks in each violin, we find that

ICLR has almost no peaks, indicating that the score distribution of the ICLR submissions is the most uniform. More details about the violin plot are explained in App. A.

## 3 Experiment

### 3.1 Experimental Setup

To demonstrate the versatility of our dataset, we conduct experiments on four tasks related to paper peer review, including acceptance prediction, score prediction, review generation, and rebuttal-discussion conversation. For the Re<sup>2</sup>-Review dataset, we sample 1,000 papers along with their reviews as the test set, with the remaining data used for training. For the Re<sup>2</sup>-Rebuttal dataset, 500 papers and the rebuttals are selected for testing, and the rest are used for training. We employ several open-source LLMs to evaluate their performance on our datasets, including LLaMA-3.1-8B-Instruct [Meta, 2024] and Qwen2.5-7B-Instruct [Yang et al., 2024]. The LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct are fine-tuned on our training data using LoRA [Hu et al., 2022] for one epoch on a learning rate of  $1 \times 10^{-4}$  with cosine scheduler, and we report both the fine-tuned and zero-shot results. In addition, for the prediction and review generation tasks, we further conduct experiments with other models specifically designed for peer review scenarios, including SEA-E [Yu et al., 2024], LLaMA-OpenReviewer-8B [Idahl and Ahmadi, 2024], DeepReviewer-14B [Zhu et al., 2025], and CycleReviewer-8B [Weng et al., 2024] (details in App. B). All the training and evaluation are conducted on four NVIDIA A100 80G GPUs. The details of the four tasks are as follows:

**Acceptance and Score Prediction.** The acceptance prediction task aims to predict whether a paper will be accepted or rejected based on its content. It is a two-class classification task (acceptance or rejection), so we use accuracy, precision, recall, and F1 score as the evaluation metrics of the acceptance prediction task. Further, score prediction focuses not just on acceptance outcomes, but on predicting the detailed review scores (e.g., overall rating) that a paper would receive. Its performance is typically measured using Mean Absolute Error (MAE) and Mean Squared Error (MSE).

**Review Generation.** The review generation task is to automatically generate peer reviews for the given paper content, simulating the feedback provided by human reviewers. Common evaluation metrics for this task include BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004], which assess the lexical overlap between generated reviews and reference reviews. In addition to the traditional metrics above, we also employ two embedding-based metrics, including the BERTScore and the cosine similarity of embedding (EmbedCos). Specifically, we adopt a 12-layer DeBERTa-large-MNLI [He et al., 2020] for BERTScore and the `sentence-transformers/all-mpnet-base-v2` model for EmbedCos. More details of these evaluation metrics are given in App. C.

**Rebuttal-Discussion Conversation.** This task aims to simulate reviewer-author interactions during the rebuttal stage of peer review, providing coherent, context-aware, and constructive feedback across turns. The basic metrics we use include BLEU, ROUGE, BERTScore, and embedding similarity. Also, to more deeply and thoroughly evaluate the ability of language models to simulate reviewer-author discussions, we further employ the LLaMA-3.1-8B-Instruct as a judge to evaluate model responses across five specific dimensions, including accuracy, constructiveness, completeness, clarity, and quality. More explanations about these five aspects and the judge instruction are given in App. D.

### 3.2 Experimental Results

**Acceptance and Score Prediction.** As shown in Tab. 2, SEA-E and DeepReviewer-14B achieve strong results in accuracy, recall, and F1 score, indicating accurate judgment of paper quality and acceptance boundaries. In contrast, LLaMA-OpenReviewer-8B and CycleReviewer-8B have higher precision but lower F1, suggesting they are more strict and conservative in accepting papers. For the open-source LLMs, LLaMA-3.1-8B and Qwen2.5-7B accept all papers without discrimination, reflecting a common flaw in LLMs: a tendency to please humans without criticism. Therefore, in the table we color them grey to indicate that the result is useless, and exclude them from the bold and underline marking of the first and second place. After finetuning, both models show more reasonable outputs, with notable decreases in MAE and MSE for score prediction, demonstrating that the finetuning on our training data significantly improves the review ability. These changes highlight the effectiveness of our dataset in enhancing models’ peer-review capabilities, enabling them to better capture the judgment basis and patterns of human reviewers.

**Review Generation.** As shown in Tab. 3, the fine-tuned LLaMA-3.1-8B ranks first among all models in both BLEU and ROUGE-L metrics, greatly outperforming its zero-shot version, indicating that the fine-tuning on our dataset allows LLaMA-3.1-8B to better match the language structure and phrasing typical of real-world peer reviews. In contrast, the results of DeepReviewer-14B and CycleReviewer-8B are obviously low on BLEU and ROUGE-L, because their generated content emphasizes abstract review reasoning rather than direct response. In terms of the EmbedCos, which evaluates similarity between semantic vectors, fine-tuned LLaMA-3.1-8B and Qwen2.5-7B achieve substantial improvements of 49.9% and 20.5% compared with the third place, respectively, suggesting a high semantic alignment between their generated texts and real reviews in the embedding space. Therefore, the training on our dataset significantly boosts models’ review generation capabilities, laying a solid foundation for building language model-based peer review assistants.

Table 2: Results comparison between open-source LLMs and baselines on prediction tasks.

Model	Metrics		Acceptance Prediction			Score Prediction	
	Accuracy	Precision	Recall	F1	MAE	MSE	
LLaMA-3.1-8B (zero-shot)	60.19±0.00	60.19±0.00	100.00±0.00	75.13±0.00	1.961±0.055	5.488±0.287	
Qwen2.5-7B (zero-shot)	60.19±0.00	60.19±0.00	100.00±0.00	75.13±0.00	2.043±0.000	5.756±0.000	
LLaMA-3.1-8B (SFT)	63.23±0.33	<b>73.66</b> ±0.64	63.74±0.72	68.35±0.12	1.141±0.030	2.290±0.108	
Qwen2.5-7B (SFT)	62.01±0.00	72.25±0.01	62.18±0.00	66.83±0.00	1.182±0.000	2.374±0.000	
SEA-E	66.24±1.42	66.69±0.68	<b>91.71</b> ±0.82	<b>77.23</b> ±0.74	1.157±0.044	2.304±0.176	
LLaMA-OpenReviewer-8B	59.76±0.21	70.92±0.25	53.59±1.05	61.05±0.77	1.197±0.012	2.413±0.021	
DeepReviewer-14B	<b>67.89</b> ±1.99	65.75±1.31	84.43±1.87	73.93±1.51	<b>1.104</b> ±0.021	<b>2.018</b> ±0.087	
CycleReviewer-8B	54.18±0.27	73.44±1.96	33.25±2.01	45.73±1.50	1.321±0.008	2.941±0.015	

Table 3: Results comparison between open-source LLMs and baselines on review generation.

Model	Metrics		BERTScore		EmbedCos
	BLEU	ROUGE-L	Precision	Recall	
LLaMA-3.1-8B (zero-shot)	1.52±0.02	16.29±0.02	59.76±0.03	58.60±0.01	0.460±0.002
Qwen2.5-7B (zero-shot)	1.37±0.00	16.17±0.00	<b>60.08</b> ±0.00	<b>60.21</b> ±0.00	0.451±0.000
LLaMA-3.1-8B (SFT)	<b>2.50</b> ±0.18	<b>17.92</b> ±0.52	59.88±0.33	53.05±0.42	<b>0.730</b> ±0.094
Qwen2.5-7B (SFT)	1.96±0.01	17.25±0.02	58.34±0.02	52.13±0.01	0.587±0.002
SEA-E	1.25±0.19	15.72±0.61	59.83±0.67	57.87±0.56	0.385±0.101
LLaMA-OpenReviewer-8B	1.49±0.05	15.90±0.31	59.38±0.48	54.39±0.60	0.444±0.008
DeepReviewer-14B	0.62±0.12	8.74±0.30	53.64±0.54	57.77±0.78	0.354±0.077
CycleReviewer-8B	0.68±0.01	11.93±0.02	54.97±0.12	53.84±0.14	0.487±0.024

**Rebuttal-Discussion Conversation.** Tab. 4 and Tab. 5 show the experimental results for the rebuttal-discussion conversation task using two categories of evaluation metrics: semantic similarity and LLM-as-judge scores. For the lexical and semantic similarity metrics (Tab. 4), the fine-tuned LLaMA-3.1-8B model achieves the best performance on four of the five metrics. Moreover, both the fine-tuned LLaMA-3.1-8B and Qwen2.5-7B consistently outperform their zero-shot versions on nearly all metrics, indicating that training on our dataset effectively enhances the LLM’s multi-turn conversation ability in the rebuttal scenario. As shown in Tab. 5, similar trends are observed in the LLM-as-judge evaluation. Across the five evaluation aspects judged by the LLM, the fine-tuned LLaMA-3.1-8B also achieves the best overall performance, and the fine-tuned two models generally surpass their zero-shot versions. These results clearly demonstrate the effectiveness of our training data in improving the capabilities of language models in author-reviewer discussion scenarios.

Table 4: Results on the lexical and semantic similarity metrics for rebuttal-discussion conversation.

Model	Metrics		BERTScore		EmbedCos
	BLEU	ROUGE-L	Precision	Recall	
LLaMA-3.1-8B (zero-shot)	1.23±0.01	13.13±0.02	52.86±0.03	<b>59.72</b> ±0.02	0.516±0.001
Qwen2.5-7B (zero-shot)	1.05±0.00	11.74±0.00	48.96±0.00	58.42±0.00	0.454±0.000
LLaMA-3.1-8B (SFT)	<b>2.07</b> ±0.32	<b>14.78</b> ±0.61	<b>54.15</b> ±0.47	58.34±0.54	<b>0.889</b> ±0.097
Qwen2.5-7B (SFT)	1.46±0.01	12.36±0.02	49.38±0.02	58.09±0.01	0.612±0.003



Table 5: Results on the LLM-as-judge metrics for rebuttal-discussion conversation.

Model \ Metrics	Quality	Constructiveness	Accuracy	Completeness	Clarity
LLaMA-3.1-8B (zero-shot)	6.936±0.013	8.380±0.013	8.032±0.003	7.133±0.005	8.297±0.011
Qwen2.5-7B (zero-shot)	6.861±0.000	8.373±0.000	7.677±0.000	6.990±0.000	7.944±0.000
LLaMA-3.1-8B (SFT)	<b>7.347</b> ±0.022	<b>8.603</b> ±0.014	<b>8.229</b> ±0.003	<b>7.210</b> ±0.006	<b>8.322</b> ±0.009
Qwen2.5-7B (SFT)	<u>7.228</u> ±0.001	<u>8.521</u> ±0.002	7.798±0.002	<u>7.183</u> ±0.001	7.936±0.001

## 4 Related Work

**Static Review Datasets.** Initially, datasets in the field of automated paper review primarily collected static data such as manuscript drafts and review comments for review process analysis, acceptance rate prediction, and review generation tasks. Kang et al. [2018] present the first public dataset of scientific peer reviews available for research purposes, and also propose two NLP tasks, acceptance prediction and aspect scores prediction based on their dataset. Dycke et al. [2022] introduce an ethically sourced multi-domain corpus of more than 5k papers and 11k review reports from five different venues. Zhang et al. [2022] conduct a thorough and rigorous study on fairness disparities in peer review with the help of LLMs, observing that the level of disparity varies and textual features are essential in reducing biases in predictive modeling. Gao et al. [2024] propose an efficient two-stage review generation framework and generate a large-scale review dataset that annotated with aspect prompts. Zhou et al. [2024] first evaluate GPT-3.5 and GPT-4 on the score prediction task and the review generation task. They also propose a dataset comprising 197 review-revision multiple-choice questions (RR-MCQ) with detailed labels from the review-rebuttal forum in ICLR 2023. Zhu et al. [2025] introduce DeepReview, a multi-stage framework designed to emulate expert reviewers by incorporating structured analysis, literature retrieval, and evidence-based argumentation. However, although the above works have made progress in scholarly peer review, most of these datasets could not guarantee the consistency between their paper content and those under review, nor did they include rebuttals, making it difficult to support the research about the reviewers-authors discussions.

**Rebuttal-included Datasets.** Different from the aforementioned works, some other researchers begin to incorporate information from the rebuttal stage into peer review datasets. Kennard et al. [2021] synthesize label sets from prior work and extend them to include fine-grained annotation of the rebuttal sentences, characterizing their context in the review and the authors’ stance towards review arguments. Jin et al. [2024] introduce AgentReview, the first LLM-based peer review simulation framework, which effectively disentangles the impacts of multiple latent factors and addresses the privacy issue. Wu et al. [2022] present a novel generation model that is capable of explicitly modeling the complicated argumentation structure from not only arguments between the reviewers and the authors, but also the inter-reviewer discussions. However, although these datasets include rebuttal content, they still suffer from several limitations, such as the issue of original manuscript versions, limited data diversity and scale, or the unreliability of simulated data.

To address the limitations mentioned above, our Re<sup>2</sup> dataset introduces several key advancements. First, we make sure that all the 19,926 papers in our dataset are their initial submission versions, ensuring version reliability and data consistency. Second, our dataset is the largest real-world peer review dataset, which features the widest range of conferences and the most complete review stages. Finally, beyond the trivial static review paradigm, we propose to treat the rebuttal and discussion data as a multi-turn conversation task between reviewers and authors, paving the way for the training and evaluation of dynamic, interactive LLM-based reviewing assistants.

## 5 Conclusion

In this work, we present Re<sup>2</sup>, the largest real-world peer review dataset that ensures data consistency, to support the model training and evaluation on tasks related to both review and rebuttal. Our Re<sup>2</sup> dataset facilitates not only traditional static tasks, but also the training of interactive, dialogue-based models for further rebuttal and discussion. In this way, not only can the review burden on reviewers be reduced, but authors are also better able to self-evaluate and improve their manuscripts before submission, ultimately helping to ease the overall reviewing pressure in the AI research community.

This not only reduces the review burden on reviewers, but also helps authors self-evaluate and improve their manuscripts before submission, easing the overall review pressure in the AI community.

**Limitations.** Although our Re<sup>2</sup> dataset represents the most comprehensive resource in the peer review field, since existing review works mainly focus on textual content, our experiments are limited to the textual and tabular components of papers, excluding visual elements such as figures (even though they are included in the released dataset). As the field advances, our future work will benefit from the integration of vision-language models, which will offer richer semantics combining textual and visual modalities and ultimately support wider applications in automated academic review.

## References

- ICLR. Assisting iclr 2025 reviewers with feedback. <https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/>, 2025.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*, 2018.
- Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. Peerasist: 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*. Springer, 2021.
- Nils Dycke, Iliia Kuznetsov, and Iryna Gurevych. Nlpeer: A unified resource for the computational study of peer review. *arXiv preprint arXiv:2211.06651*, 2022.
- Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. Mred: A meta-review dataset for structure-controllable text generation. *arXiv preprint arXiv:2110.07474*, 2021.
- Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv preprint arXiv:2211.06398*, 2022.
- Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.
- Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022.
- Mike D’Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*, 2023.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*, 2024.
- Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*, 2024.
- Ruiyang Zhou, Lu Chen, and Kai Yu. 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)*, 2024.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cyclereviewer: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.
- Neha Kennard, Tim O’Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. Disapere: A dataset for discourse structure in peer review discussions. *arXiv preprint arXiv:2110.08520*, 2021.
- Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. Exploring jiu-jitsu argumentation for writing peer review rebuttals. *arXiv preprint arXiv:2311.03998*, 2023.
- Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. Politepeer: does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation*, 2024.

- Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Meta. Introducing llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>, 2024.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.
- Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *arXiv preprint arXiv:2407.12857*, 2024.
- Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. *arXiv preprint arXiv:2412.11948*, 2024.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

## A Details of the Violin Plot

To illustrate the (normalized) review score distributions of each conference, we utilize the violin plot to conduct a statistical analysis of the normalized scores in Fig. 2(f). Here we introduce more details of the violin plot. The violin plot is a statistical graphic that combines the box plot and the kernel density plot, allowing for a better display of the data distribution. As shown in Fig. 3, each violin represents the score distribution of a conference.

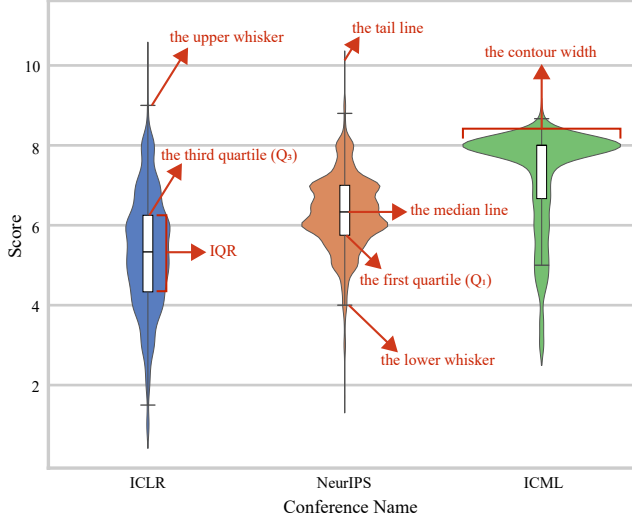


Figure 3: **Details of the Violin Plot.** The upper and lower whiskers represent the maximum and minimum observed values. The third quartile ( $Q_3$ ) is the value below which 75% of the data falls. The first quartile ( $Q_1$ ) is the value below which 25% of the data falls. IQR represents the distribution of the middle 50% of the data. The tail line can extend beyond the data boundaries, reflecting the smoothness of the curve. The median line’s position varies with the data distribution. The contour width represents the density of the number of papers.

**Kernel density estimation.** The wider the contour, the more papers are assigned this score. The smoothness of the curve is controlled by the bandwidth. The larger the value, the smoother the curve. This also results in tail lines that go beyond the data boundaries, with no real data points typically present in them. The kernel density estimation function is as follows:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (1)$$

where  $\hat{f}(x)$  denotes the estimated density function value at position  $x$ ;  $n$  denotes the sample size;  $h$  denotes the bandwidth, which controls the smoothness of the kernel function;  $x_i$  denotes the  $i$ -th sample point;  $K(\cdot)$  denotes the kernel function, usually a symmetric probability density function. We use Seaborn<sup>2</sup> to plot the violin plot, which defaults to the Gaussian kernel.  $x$  represents the estimation point and  $x_i$  represents the true sample point. The kernel function is as follows:

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - x_i}{h}\right)^2\right) \quad (2)$$

**Boxplot.** The box plot consists of the box, the median line, and the upper and lower whiskers. The upper boundary of the box is the third quartile ( $Q_3$ ), which indicates that 75% of the data points are below this value; the lower boundary of the box is the first quartile ( $Q_1$ ), which indicates that 25% of the data points are below this value. The total height of the box is the Interquartile Range ( $IQR = Q_3 - Q_1$ ), which represents the spread of the middle 50% of the data.

<sup>2</sup><https://seaborn.pydata.org/>

The median line ( $Q_2$ ) is usually located in the center of the box, but if the data distribution is not balanced, the position of the line may shift. When most of the data is concentrated in the lower values, the median line will be closer to the lower boundary; when most of the data is concentrated in the higher values, the median line will be closer to the upper boundary. In Fig. 3, for ICML, the median line is near the third quartile because most of the papers have scores around 8.

The upper and lower whiskers represent the maximum and minimum observed values, and they are typically used to identify outliers. The formulas are as follows:

$$\text{Lower Whisker} = \max(\min(X), Q_1 - 1.5 \times \text{IQR}) \quad (3)$$

$$\text{Upper Whisker} = \min(\max(X), Q_3 + 1.5 \times \text{IQR}) \quad (4)$$

When analyzing the distribution of review scores, most of the data falls between the upper and lower whiskers. As shown in equations 3 and 4, neither the upper whisker nor the lower whisker represents the actual maximum or minimum values. Since very few papers have such low scores, these scores can be considered outliers and placed below the lower whisker. From Fig. 3, we can see that the curve for ICLR is smoother compared to the other two conferences, with fewer peaks. Therefore, the paper score distribution for ICLR is the most evenly spread.

## B Details of the Peer Review Methods in the Experiment

In addition to the open-source LLMs like LLaMA-3.1-8B-Inst and Qwen2.5-7B-Inst, we further evaluate some models specifically designed for paper review scenarios on our test dataset, including SEA-E [Yu et al., 2024], LLaMA-OpenReviewer-8B [Idahl and Ahmadi, 2024], DeepReviewer-14B [Zhu et al., 2025], and CycleReviewer-8B [Weng et al., 2024]. The details of these methods are given below:

- SEA-E [Yu et al., 2024]: It is the evaluation model from an automated paper reviewing framework named SEA, which comprises of three modules: Standardization, Evaluation, and Analysis. SEA-E utilizes the standardized data that is integrated from multiple reviews for fine-tuning, enabling it to generate constructive reviews.
- LLaMA-OpenReviewer-8B [Idahl and Ahmadi, 2024]: The OpenReviewer is an open-source system for generating high-quality peer reviews of machine learning and AI conference papers. Its core is LLaMA-OpenReviewer-8B, an 8B parameter language model specifically fine-tuned on 79,000 expert reviews from top conferences, which produces considerably more critical and realistic reviews compared to general-purpose LLMs like GPT-4 and Claude-3.5.
- DeepReviewer-14B [Zhu et al., 2025]: The DeepReview is a multi-stage framework designed to emulate expert reviewers by incorporating structured analysis, literature retrieval, and evidence-based argumentation. Using DeepReview-13K, a curated dataset with structured annotations, DeepReviewer-14B is trained and outperforms CycleReviewer-70B with fewer tokens. In its best mode, DeepReviewer-14B achieves win rates of 88.21% and 80.20% against GPT-o1 and DeepSeek-R1 in evaluations.
- CycleReviewer-8B [Weng et al., 2024]: The author explores the feasibility of using open-source post-trained LLMs as autonomous agents capable of performing the full cycle of automated research and review, from literature review and manuscript preparation to peer review and paper refinement. In this iterative preference training framework, CycleReviewer simulates the peer review process, providing iterative feedback via reinforcement learning.

## C Details of the Evaluation Metrics

**BLEU.** BLEU (Bilingual Evaluation Understudy) mainly measures the similarity between texts using n-gram overlap. An n-gram is a sequence of  $n$  consecutive words or characters in a text. The n-gram overlap refers to how many n-grams in the generated text match with those in the reference text, reflecting local similarity. For n-grams, a smaller  $n$  makes it easier to match, but it loses context; a larger  $n$  makes it harder to match, but better reflects sentence structure and includes more contextual information. Therefore, we need to combine the n-gram precision for different values of  $n$ , usually

by calculating the geometric mean, and add a penalty factor (BP) to prevent the generated text from being too short. The formulas are as follows:

$$P_n = \frac{\sum_{C \in \text{Candidates}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C \in \text{Candidates}} \sum_{\text{n-gram} \in C} \text{Count}(\text{n-gram})} \quad (5)$$

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases} \quad (6)$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (7)$$

Here BP stands for brevity penalty, which is used to penalize overly short outputs;  $c$  denotes the length of the generated text;  $r$  denotes the length of the reference text;  $w_n$  denotes the weight of the  $n$ -gram (usually  $\frac{1}{n}$ ). It’s important to note that BLEU is not sensitive to synonyms or grammatical variations. It doesn’t work well for evaluating single sentences and is more suitable for evaluating longer texts or entire documents, which matches the tasks about review generation we focus on.

**ROUGE.** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is commonly used to evaluate how much of the reference text is covered by the output from tasks like text summarization and question answering. It has several variants, such as ROUGE-N (based on  $n$ -gram recall), ROUGE-L (based on the Longest Common Subsequence, or LCS), ROUGE-W (weighted LCS), and ROUGE-S / ROUGE-SU (based on skip-bigram matches).

Taking ROUGE-L as an example, the LCS is the longest sequence of words that appear in both texts in the same order, but not necessarily next to each other. In our experiment, we report the ROUGE-L F1 score, which is the harmonic mean of ROUGE-L Precision and ROUGE-L Recall.

$$\text{ROUGE-L}_{\text{Precision}} = \frac{L}{m}, \quad \text{ROUGE-L}_{\text{Recall}} = \frac{L}{n} \quad (8)$$

$$\text{ROUGE-L}_{F_\beta} = \frac{(1 + \beta^2) \cdot \text{ROUGE-L}_{\text{Precision}} \cdot \text{ROUGE-L}_{\text{Recall}}}{\beta^2 \cdot \text{ROUGE-L}_{\text{Precision}} + \text{ROUGE-L}_{\text{Recall}}} \quad (9)$$

Here  $L$  is the length of the LCS,  $m$  is the length of the reference text, and  $n$  is the length of the generated text. When there are multiple reference texts, ROUGE compares the generated text with each one, calculates a ROUGE-L score for each pair, and then selects the highest ROUGE-L score as the final result.

ROUGE has similar limitations to BLEU. Since both are based on lexical matching at the surface level, they cannot effectively evaluate semantic similarity. Therefore, we often further use semantic evaluation metrics, such as BERTScore and the cosine similarity of embeddings (EmbedCos).

**BERTScore.** BERTScore uses BERT to extract word embeddings and measures how well two texts match by comparing the similarity between their word vectors. First, BERT embeddings are generated separately for  $r_i$  (the reference) and  $c_i$  (the generated text). Then, using the word embeddings  $r_i$  and  $c_i$ , pairwise cosine similarities are calculated to form a similarity matrix  $S_{ij}$  of size  $n \times m$ :

$$S_{ij} = \cos(\vec{c}_i, \vec{r}_j) = \frac{\vec{c}_i \cdot \vec{r}_j}{\|\vec{c}_i\| \cdot \|\vec{r}_j\|} \quad (10)$$

After obtaining  $S_{ij}$ , we can calculate the Precision and Recall of BERTScore. Precision is computed by finding, for each word in the candidate sentence, the most similar word in the reference sentence, and then taking the average of these maximum similarities. Recall is calculated by doing the same in reverse: for each word in the reference sentence, find the most similar word in the candidate sentence, and then take the average.

$$\text{Precision} = \frac{1}{n} \sum_{i=1}^n \max_{1 \leq j \leq m} S_{ij}, \quad \text{Recall} = \frac{1}{m} \sum_{j=1}^m \max_{1 \leq i \leq n} S_{ij} \quad (11)$$

BERTScore does not require exact word matching, allowing it to recognize synonyms to some extent and providing some level of contextual awareness. However, it is entirely based on word-level matching and does not take sentence structure or grammatical order into account. It requires a large

amount of computation, and its results are harder to interpret compared to BLEU and ROUGE. In our work, we use a 12-layer DeBERTa-large-MNLI for the strong semantic understanding capability.

**EmbedCos.** EmbedCos refers to the cosine similarity of embeddings. We use the `sentence-transformers/all-mpnet-base-v2` model to calculate EmbedCos, because compared to other base Transformer models like BERT and RoBERTa, MPNet is more efficient and can generate high-quality sentence embeddings in less time. This model focuses on sentence-level embeddings and optimizes the semantic distance between sentences, so similar sentences are closer together in the vector space, and different sentences are farther apart.

The calculation process of EmbedCos can be divided into two steps: First, a set of sentences is tokenized and passed into MPNet. The embeddings obtained here represent the entire sentence’s semantic meaning, unlike BERTScore, which uses individual words. Second, based on the obtained sentence embeddings, the EmbedCos is computed as the cosine similarity between sentences:

$$\text{EmbedCos}(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \cdot \|\vec{v}\|} \in [-1, 1] \quad (12)$$

Here  $\vec{u} \cdot \vec{v}$  is the dot product of the vectors, and  $\|\vec{u}\|$  is the L2 norm of  $\vec{u}$ .

As shown in Tab. 6, we show the comparison between the four metrics: BLEU, ROUGE-L, BERTScore, and EmbedCos.

Table 6: Comparison of the evaluation metrics.

Metric	BLEU	ROUGE-L	BERTScore	EmbedCos
<b>Evaluation Method</b>	N-gram Precision	LCS	Token-level Semantic Similarity (BERT)	Cosine Similarity of Sentence Embeddings
<b>Semantic Focus</b>			✓	✓
<b>Model Dependency</b>			✓	✓
<b>Evaluation Granularity</b>	N-gram	LCS	Token-level Similarity Aggregation	Sentence-level Embedding Similarity

## D Details of the LLM as Judge

In the rebuttal-discussion conversation task, we use LLaMA-3.1-8B-Inst to evaluate the quality of generated dialogue statements. Here we instruct the judge to focus on five aspects to evaluate the results generated by the reviewer model, including accuracy, clarity, constructiveness, completeness, and quality.

- Accuracy is selected to ensure the correctness of the model’s output, making sure the generated responses align logically with the input questions or context and avoid factual errors.
- Clarity assesses whether the generated statements are expressed fluently and understandably, avoiding ambiguity or redundancy, thereby reflecting the model’s language organization capability.
- Constructiveness focuses on the relevance and practical value of the generated content, such as whether it can provide effective improvement suggestions for contentious points, demonstrating the usefulness of the dialogue.
- Completeness measures whether the response covers key issues or arguments, avoiding the omission of important information, especially in academic rebuttals where comprehensive responses to critiques are essential.
- Quality serves as an overall evaluation criterion, integrating all the above dimensions to grade the generated content.



Among these metrics, accuracy and clarity emphasize the model's output quality at a technical level; constructiveness emphasizes dialogue relevance at a practical level; completeness focuses on comment coverage at a content level; and quality acts as a global indicator.

The instruction we used to guide the LLaMA-3.1-8B-Inst as a judge is shown as below:

```
You are an expert evaluator.
Given the gold reference answer and a candidate answer, score the
candidate's quality, constructive, accuracy, completeness and
clarity on a scale of 1-10.
- Quality: Scores overall depth, logic, and usefulness-low for
shallow/chaotic content, high for insightful and valuable input.
- Constructive: Measures whether the comment offers solutions or
fosters discussion-low for pure criticism, high for actionable
feedback.
- Accuracy: Rates factual/logical correctness-low for
errors/misleading claims, high for well-supported and precise
statements.
- Completeness: Assesses coverage of key points-low for major
omissions, high for thorough and detailed analysis.
- Clarity: Judges coherence and readability-low for
confusing/verbose language, high for concise and well-structured
delivery.
Please apply stricter grading criteria to reduce the proportion
of high scores and ensure a reasonable score distribution. Only
exceptionally outstanding performances should receive high scores
, average performances should receive moderate scores, and poor
performances should receive low scores. Be more objective and
conservative in your grading.
Respond strictly as JSON, do not provide any other content:
{
  "quality": <int>,
  "constructive": <int>,
  "accuracy": <int>,
  "completeness": <int>,
  "clarity": <int>
}
```