An Analysis of Tasks and Datasets in Peer Reviewing

Moritz Staudinger[1] **Wojciech Kusa**[1,2] **Florina Piroi**[1] **Allan Hanbury**[1]
[1]TU Wien, Vienna, Austria
[2]Allegro, Warsaw, Poland
{moritz.staudinger,wojciech.kusa,florina.piroi,allan.hanbury}@tuwien.ac.at

Abstract

Taking note of the current challenges of the peer review system, this paper inventories the research tasks for analysing and possibly automating parts of the reviewing process, like matching submissions with a reviewer's domain of expertise. For each of these tasks we list their associated datasets, analysing their quality in terms of available documentation of creation and use. Building up on this, we give a set of recommendations to take into account when collecting and releasing data.

1 Introduction

Peer Reviewing is a vital part of academic integrity since as early as 1665, when Henry Oldenburg founded the Philosophical Transactions of the Royal Society and established the concept of reviewing the work of others (Kachooei and Ebrahimzadeh, 2022). Since then, the concept of peer review has remained largely unchanged, and nearly every journal has adopted it to select the best publications to publish. In recent decades, peer reviewing has become more of a burden for various reasons, like the overwhelming number of article submissions to review, their often less than optimal quality of content, lack of time to keep up with the majority of published research, or the fact that reviewing is voluntary and researchers find they have less time to review other work.

Technological advances that allow to share research and cooperate worldwide combined with the increasing number of researchers have caused a dramatic increase in the research published daily. The availability of Large Language Models (LLMs) has in fact simplified some tasks of the scientific writing process, like proof reading or text reformulation. This means that the time from idea to written research is significantly shorter. In the Computer Science domain, the number of publications over the last 10 years (2014–2023) has increased dramatically, submissions to ACL have increased sevenfold, and submissions to NeurIPS eight-fold[1].

Accounting for these factors, it is harder to find sufficient numbers of reviewers for such conferences. In 2016, about 33% of the NeurIPS reviewers were PhD students (Shah et al., 2018), a percentage that is most likely higher today. The outcome is that there are more inexperienced researchers that have to review larger numbers of submissions, eventually leading to reviews of lower quality.

In recent years, there are efforts to automate and optimize the different peer review steps to lighten the load of the reviewers, area chairs, and programme committees[2]. Researchers interested in this domain have defined research tasks and collected data to address them.

In this work, we make an inventory of tasks and datasets associated with Peer Review Automation and Optimization, and give recommendations on how to collect and publish such datasets. Without claiming that our work is exhaustive, we see it as a necessary contribution to this field. We examine the available datasets, the tasks they have been used or created for, discussing use cases and automation tasks.This work complements the very recently published Dagstuhl Proceedings (Kuznetsov et al., 2024), and the analyses in Drori and Te'eni (2024) or Lin et al. (2023a).

The paper is structured as follows: Section 2 gives an account of the tasks and their datasets in the Peer Review Analysis domain. Section 3 shows the overlaps between various datasets. Section 4 discusses some recommendations on the creation of datasets in the Peer Review domain. Section 5 comments on the current advances in this domain, concluding with Section 6.

[1] https://github.com/lixin4ever/Conference-Acceptance-Rate, (Access: May 2024)

[2] A description of these roles is available at https://aclrollingreview.org/reviewing

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 257–268

2 Datasets and Tasks

In our effort to understand the research challenges in the scientific peer reviewing domain we analyzed 53 datasets that contain peer reviews. For each of them, mostly created in the last 5 years, we looked at the research tasks for which these sets were used.

Datasets: Of these datasets, 37 cover publications that are in the ML and NLP scientific domains, obtained from the OpenReview portal[3], 6 are in the NLP domain and obtained from ACL Rolling Review[4], 5 datasets cover multiple domains and journals, 3 obtained from F1000[5] and 2 from PeerJ[6]. In Tables 1 to 8 we use the following abbreviations to indicate the source of the collected reviews: OR (OpenReview), ACL (ACL Rolling Review), CO (CoNLL and COLING), F (F1000 Journal), PJ (PeerJ), RS (ShitMyReviewerSays), and SG for self gathered data.

Though most of the datasets can be used for more than one task, in our work we choose to assign them to the task they were used for at their creation time. What we observed is that most of the researchers in this domain opt to create their own datasets from scratch using the OpenReview API, instead of using one of the many available datasets. This affects the reproducibility and comparability of research results, as not all datasets are publicly available or well documented to foster their reuse.

Tasks: Reviews can be looked at with different research aims in mind. Earlier work (Cho, 2008), for example, analyzed the helpfulness of student reviews, while more recent work looks at various aspects of peer reviewing, like Discourse Analysis (Kennard et al., 2022; Ruggeri et al., 2023), Peer Review Quality (Verma et al., 2023; Ghosal et al., 2022b; Bharti et al., 2023) or meta review generation (Chen et al., 2023; Ridenour et al., 2022).

In the following we will discuss the most common tasks and datasets for peer reviewing.

2.1 Review Analysis

Tasks grouped under this title aim to understand how reviews impact the submissions they were written for, aiming to detect biases in scoring, for example. For this area we identified three datasets (Table 1). The first one listed is used to look at how review scores for submissions to ICLR 2017-2022 affect their acceptance rate and their future citations (Wang et al., 2023). The second listed dataset was used in research that looked at peer reviews for journal submissions in the medical domain (BMC, BMJ, PLOS Medicine), examining their length and quality (Geldsetzer et al., 2023). The "ICLR Database" (Zhang et al., 2022) has been used in research that analysed fairness disparities in the review process and the geographical and institutional distribution of submissions.

Dataset	Year	Public	Origin
Wang et al. (2023)	2023	Yes	OR
Geldsetzer et al. (2023)	2023	Partly	
ICLR Database (Zhang et al., 2022)	2022	Yes	OR

Table 1: Datasets used in Review Analysis tasks.

2.2 Reviewer Assignment

In content preparatory phases of a conference, submissions are assigned to reviewers, ideally, based on their experience and interests. Most often, this is done via a bidding process, after which reviewers get assigned papers to review, according to their bids. This bidding process is, nevertheless, prone to manipulation, such as reviewers either voting to review submissions from their friends and close colleagues (choosing to grade them highly) or, when having sent in submissions, bidding to review similar submissions to their own, then negatively scoring them to improve their own submission acceptance chances, as the dataset (Jecmen et al., 2023) and research in Jecmen et al. (2024) demonstrate. To avoid this kind of malicious bidding, one possibility is to automatically match submissions to reviewers. Two further datasets created to investigate the effectiveness of such matching and, at the same time, to analyse the impact of malicious bidding are listed in Table 2.

The dataset and the method introduced by Stelmakh et al. (2020) detect malicious behavior at the meta-review level. To evaluate different methods of fair assignment of reviewers, Stelmakh et al. (2023) created a gold-level dataset on the self-reported expertise of reviewers for given publications.

Dataset	Year	Public	Origin
Malicious Bidding (Jecmen et al., 2023)	2023	Yes	SG
Reviewer-Paper-Match (Stelmakh et al., 2023)	2023	Yes	SG
Stelmakh et al. (2020)	2020	Yes	SG

Table 2: Datasets used in Reviewer Assignment tasks.

[3] https://openreview.net/
[4] https://aclrollingreview.org
[5] https://f1000research.com/
[6] https://peerj.com/

2.3 Score and Acceptance Prediction

To ease decision making, reviewers are asked to numerically score the different aspects of a submission, such as *Soundness, Presentation, Contribution*, and give an *Overall Score*. These scores are then used by area chairs to decide on the acceptance of a submission, and to formulate their meta-review. Fernandes and Vaz-de Melo (2023, 2022) analyzed the *Review Score Prediction (RSP)* and the *Paper Decision Prediction (PDP)* tasks, by evaluating models on two self-mined datasets of ICLR conference submissions (Table 3). The two datasets mainly introduced in their research differ in the number of reviews (14,459 vs 18,170) in them. In their work, they trained models on reviews, to predict RSP and PDP values on Submission-Review tuples. This was similar to the dataset and prediction task, proposed by Kang et al. (2018), who created one of the first Peer Review datasets, to allow score prediction as well as review generation.

Liu et al. (2023b) conducted a study on the willingness of reviewers to change their scores after the author rebuttal and revision phases. Specifically, they conducted a randomized controlled trial on 108 participants to find under which conditions reviewers tend to change their original scores. With PEERAssist, Bharti et al. (2021) explored score prediction with a deep neural architecture, and outperformed a similar architecture using sentiment analysis (Ghosal et al., 2019). Ribeiro et al. (2021) used a private dataset of two non-disclosed conferences to perform RSP, PDP, and sentiment analysis to extract review polarities. Datasets that tackle these tasks are listed in Table 3.

Dataset	Year	Public	Origin
Fernandes and Vaz-de Melo (2023)	2023	Yes	OR
ReviewerAnchoring (Liu et al., 2023b)	2023	Yes	OR
Fernandes and Vaz-de Melo (2022)	2022	Yes	OR
PEERAssist (Bharti et al., 2021)	2021	Yes	OR
Ribeiro et al. (2021)	2021	No	
PeerRead (Kang et al., 2018)	2018	Yes	OR,ACL,CO

Table 3: Score and Acceptance Prediction datasets.

2.4 Revision, Dialogue, Rebuttal Analysis

Some peer review processes include a dialogue between the authors and the reviewers, a dialogue with visible effects on the final version of a publication. Analyzing differences between consequent submission versions as well as the dialogue among reviewers or between reviewers and authors is, therefore, of interest. Having access to such information allow experiments on predicting changes over time or finding disagreements between different parties. Table 4 lists the peer review datasets used for approaching any of these tasks.

Among the datasets that investigate the differences between submission versions we mention CASIMIR (Jourdan et al., 2024), ARIES (D'Arcy et al., 2023), and arXivEdits (Jiang et al., 2022). The CASIMIR collection contains 3.7 million sentence pairs, from different submission revisions. A possible use case is the analysis of differences between revisions and prediction necessary revisions. The ARIES dataset has been created for the task where revised submissions are annotated to trace the edits back to the reviewer comments. arXivEdits does not contain reviews, but a set of 751 arXiv publications with revision history and sentence alignment across multiple paper revisions.

Datasets were created with the aim to understand and classify the arguments and disagreements made in the reviews, like APE (Argument Pair Extraction) (Cheng et al., 2020) and the AMPERE dataset (Hua et al., 2019). The APE dataset contains 4,764 argument pairs from the review-rebuttal phases, while the AMPERE dataset contains proposition and argument type annotations, aiming to answer research questions on argument classification. In the same direction, Singh et al. (2021) analyzed peer reviews looking at discussion comparisons, to better understand the factors that affect the acceptance decisions. Another dataset that explores disagreements between reviewer comments is ContraSciView (Kumar et al., 2023), a dataset of around 28,000 annotated review pairs, if they agree or contradict each other. They further proposed a model to detect contradictory statements in Peer Reviews. Finally, here, we note the ArgSciChat dataset (Ruggeri et al., 2023) containing argumentative dialogues, argumentative and exploratory questions and answers for 20 papers in the NLP domain.

Rebuttal strategies are of interest to researchers, with data collected for and work looking into rebuttal generation based on reviewers' comments and evaluations (the Jiu-Jitsu dataset (Purkayastha et al., 2023)). Huang et al. (2023) conducted an empirical study on ICLR papers, to detect the most common rebuttal strategies and predict the changes in review scores before and after the rebuttal phase. Similarly, DISAPERE (DIscourse Structure in Academic PEer REview) (Kennard et al., 2022) is a dataset of 20,000 labeled review-rebuttal sentences

combinations, to analyze the interpretation of reviews and response strategies. ReAct (Choudhary et al., 2021) is a dataset that falls into this group (rebuttal and revision strategy support). The authors annotated review sentences such that, on one side, review recommendation can be quickly classified into actionable (possible to answer in a revision), and on the other side the revision action can be classified by finer-grained types, like *Suggestion, Question,* or *Disagreement*.

Dataset	Year	Public	Origin
CASIMIR (Jourdan et al., 2024)	2024	Yes	OR
Huang et al. (2023)	2023	No	OR
Jiu-Jitsu (Purkayastha et al., 2023)	2023	Yes	OR
ARIES (D'Arcy et al., 2023)	2023	Yes	OR
ArgSciChat (Ruggeri et al., 2023)	2023	Yes	SG
ContraSciView (Kumar et al., 2023)	2023	Yes	OR
arXivEdits (Jiang et al., 2022)	2022	Yes	
DISAPERE (Kennard et al., 2022)	2022	Yes	OR
COMPARE (Singh et al., 2021)	2021	Yes	OR
ReAct (Choudhary et al., 2021)	2021	Yes	OR
APE (Cheng et al., 2020)	2020	Yes	OR
AMPERE (Hua et al., 2019)	2019	Yes	OR, ACL

Table 4: Datasets for Revision, Dialogue and Rebuttal analysis and prediction.

2.5 Review Generation

While review generation may solve some of the reviewing issues, such as speeding up the academic review process or generating high quality reviews, it may also introduce new threats into the academic review process. For example, authors could tune their submissions so that automated review models would accept them.

Table 5 shows a list of datasets suitable for review generation tasks. The first model for automatic review generation is ReviewRobot (Wang et al., 2020), trained on data from OpenReview combined with multiple knowledge graphs to obtain domain knowledge to conduct reviews. In 2022 Yuan et al. (2022) published their ASAP-Review dataset and a model for review generation, but found that their model is less constructive and less factual then human-written reviews. Another dataset published for researchers to work on the generation of reviews and the prediction of scoring results is the ORB dataset (Szumega et al., 2023). The dataset is, though, not well documented.

With every new LLM release, researchers make use of them to generate reviews. One dataset to enable their use is, for example, the Review-Revision Multiple-Choice Questions (RR-MCQ) dataset, that allows to have a quick assessment on LLM performance in the scientific review domain (Zhou et al., 2024). Similarly, Liu and Shah (2023) created a dataset to identify errors in computer science publications, where they manually modified the publications to contain factual inaccuracies.

Kuznetsov et al. (2022) looked into assisting the peer review process providing several datasets for tasks like review generation, pragmatic tagging, or score prediction. The Yes-Yes-Yes dataset (Dycke et al., 2022) contains review from the ACL Rolling Review, 2021, and, later, the same authors created F1000RD (Kuznetsov et al., 2022), the first multidomain corpus of peer reviews of the F1000Research platform. These datasets were combined and released later (Dycke et al., 2023a). Based on the F1000RD dataset, a Pragmatic Tagging dataset was released for the PragTag Shared-Task (Dycke et al., 2023b).

Dataset	Year	Public	Origin
RR-MCQ (Zhou et al., 2024)	2024	Yes	OR
NLPEER (Dycke et al., 2023a)	2023	Yes	ACL,CO,F
PragTag (Dycke et al., 2023b)	2023	Yes	F
ORB (Szumega et al., 2023)	2023	Yes	OR
ReviewerGPT (Liu and Shah, 2023)	2023	Partly	OR
Yes-Yes-Yes (Dycke et al., 2023a)	2022	Yes	ACL
F1000RD (Kuznetsov et al., 2022)	2022	Yes	F
ASAP-Review (Yuan et al., 2022)	2022	Yes	OR
ReviewRobot (Wang et al., 2020)	2020	Yes	OR

Table 5: Datasets for Review Generation.

2.6 Metareview Generation

Writing metareviews is a task performed by Area Chairs, who read through all the reviews for a submission, check on disagreements between them, and provide a summary of the reviews with a final decision. Table 6 lists the datasets used in research on metareview analysis and generation.

Sun et al. (2024b) created an own peer review dataset, covering five years of review data from OpenReview, and trained a model on it to generate metareview texts. PeerSum (Li et al., 2023) is a different dataset for metareview generation, which features explicit conflicting information in source documents that the meta-reviewers have to handle. ORSUM (Zeng et al., 2023) is a corpus for scientific opinion summarization, which uses a undisclosed dataset from 39 conferences and workshops that published their peer reviews on OpenReview. The MOPRD (Multidsciplinary Open Peer Review Dataset) (Lin et al., 2023b) dataset was created to enable metareview generation, editorial decision prediction, author rebuttal generation, and scientometric analysis of papers from multiple disciplines.

It is composed of paper metadata, manuscripts of the initial submission and multiple revisions, review comments, meta-reviews, author rebuttal letters, and editorial decisions of papers from Biology, Chemistry, Computer Science, Environment and Medicine published on PeerJ. Wu et al. (2022) constructed a Peer Review and Rebuttal Counter-Arguments (PRRCA) dataset, which contains both the reviews, and the authors responses to model the peer review evaluation process.

The Meta-Review Dataset (MReD) (Shen et al., 2022) is a sentence-level annotated dataset of 23,675 reviews and their corresponding 7,089 meta-reviews. Each sentence is categorized into one of nine categories (abstract, strength, weakness, rating summary, area chair disagreement, rebuttal process, suggestion, decision, and miscellaneous) to then generate the meta-reviews based on these sentences.

Dataset	Year	Public	Origin
MetaWriter (Sun et al., 2024b)	2024	No	OR
PeerSum (Li et al., 2023)	2023	Yes	OR
ORSUM (Zeng et al., 2023)	2023	Yes	OR
MOPRD (Lin et al., 2023b)	2023	No	PJ
PRRCA (Wu et al., 2022)	2022	Yes	OR
MReD (Shen et al., 2022)	2022	Yes	OR

Table 6: Overview of Metareview Datasets.

2.7 Quality Analysis

Analyzing reviews in scientific papers can help to measure the helpfulness of the conducted peer review, the politeness or harshness of the peer reviewer, the reviewer confidence on a given topic, or the reviewer disagreement. Research in this area aims to detect and measure quality issues in conducted reviews. Table 7 presents a list of datasets for quality analysis of peer reviews.

The datasets in this subsection are used in, roughly seen, two types of tasks. One is classifying the reviews by detecting politeness levels and harshness (e.g., PolitePEER in Bharti et al. (2023) or Verma et al. (2022)), and detecting review polarities (Wang and Wan, 2018). The other task is exploring the content of the reviews from different argumentative points of view, in order to detect disagreement between reviewers and scoring differences between reviewers and submission versions (Chakraborty et al., 2020; Gao et al., 2019), or to estimate the review helpfulness by checking its thoroughness (Severin et al., 2023) and detecting suggestions for improvement to the submission authors (Pfütze et al., 2022), finding arguments in peer reviews as in AMSR (Argument Mining in Scientific Reviews) (Fromm et al., 2021).

A combination of the two types of tasks is seen in ReviVal (Verma et al., 2023), in which the authors use the dataset to grade reviews based on *Exhaustiveness* and *Strength* of the review texts, or in SubstanReview (Guo et al., 2023), in which the authors look at substantiation levels of reviews by verifying that subjective statements (claims) are backed up by justifications (evidence). HedgePeer (Ghosal et al., 2022b) provides a dataset for uncertainty detection in peer reviews by annotating the review comments in terms of hedge cues and spans. Finally, the dataset introduced in by Ghosal et al. (2022a) aims to facilitate the analysis of the reviewers' confidence on certain sections and aspects of the paper and then use those insights to investigate further the quality of peer reviews.

Dataset	Year	Public	Origin
PolitePEER (Bharti et al., 2023)	2023	Yes	OR,RS
ReviVal (Verma et al., 2023)	2023	No	OR
SubstanReview (Guo et al., 2023)	2023	Yes	ACL,CO
Severin et al. (2023)	2023	Yes	
HedgePeer (Ghosal et al., 2022b)	2022	Yes	OR
Ghosal et al. (2022a)	2022	Yes	OR
Pfütze et al. (2022)	2022	No	SG
Verma et al. (2022)	2022	Yes	OR,RS
AMSR(Fromm et al., 2021)	2021	Yes	OR
Chakraborty et al. (2020)	2020	No	OR
ACL-2018(Gao et al., 2019)	2019	Yes	ACL
Wang and Wan (2018)	2018	No	OR

Table 7: Review Quality Analysis datasets.

2.8 Citation Prediction

An interesting task, which is not directly associated with the Peer Review domain, is to predict the popularity and citations of accepted publications based on their reviews. To our knowledge, there currently are two datasets on this topic (see Table 8). The CiteTracked (Plank and Dalen, 2019) dataset links reviews from the NeurIPS conference with their citations to predict the impact of the submissions. Li et al. (2019) performed a similar study on ICLR reviews, but provided further data and metadata to their models, such as information about the authors.

Dataset	Year	Public	Origin
CiteTracked (Plank and Dalen, 2019)	2019	No	OR
Citation-Count-Pred (Li et al., 2019)	2019	Yes	OR

Table 8: Citation Prediction Datasets.

3 Dataset Similarities and Discrepancies

In the last years, OpenReview has become the primary data source for review-related tasks, and, therefore, many of the published datasets overlap. We look more closely into the datasets which, among others, contain the ICLR conference review data (35 sets, see Table 9).

We detected several inconsistencies across these datasets coming from the same primary source (here, OpenReview). One of them is in the number of reviews for one year of the conference (2017). Namely, Kang et al. (2018); Verma et al. (2023) report 427 submissions, while Wang and Wan (2018); Zhang et al. (2022) and Chakraborty et al. (2020) reported 490 and 491 submissions (the latter difference possibly being a reporting error on the workshop track). Later, Bharti et al. (2021) cleaned this dataset by removing duplicates or empty reviews and reports 354 qualitative reviews for ICLR 2017. These differences can be due to authors agreeing to publish their submissions later than the dataset creation time, not filtering out duplicates, including workshop submissions[7], or, more likely, are due to reporting inaccuracies. We found, for example, that Zhang et al. (2022) reported 47 workshop papers. Some studies excluded data due to poor quality or availability issues (Fernandes and Vaz-de Melo, 2023; Cheng et al., 2020; Chakraborty et al., 2020), while other studies avoid any kind of reporting on data quality and filtering.

Most of the datasets were extended with annotations to allow the evaluation of the conducted experiments. However, the number of annotations is rather small (in the hundreds), and do not follow a common structure, to allow the integration of multiple datasets and the interchangeability of datasets. While many of the datasets contain several hundred annotated sentences or reviews, such as D'Arcy et al. (2023) and Zhou et al. (2024); others, like Kumar et al. (2023) and Jourdan et al. (2024) annotated larger amounts of data. The ASAP-Review dataset was extended by manually annotating 8,582 publications with sentence level labels on sentiment and aspect categories (Kumar et al., 2023; Yuan et al., 2022). Jourdan et al. (2024) automatically aligned more than 3.7 million sentence pairs for revision analysis and labeled the most likely intention behind each edit.

[7]Workshop submissions are different from the main conference submissions as they are often reviewed more leniently than conference submissions.

Furthermore, different conferences use different review guidelines and numerical scoring scales, which can lead to biases if these conferences are combined. For example, ICTIR allows reviewers to score papers on a scale from -3 to 3, while NeurIPS uses a scale from 1 to 10.

4 Guidelines for Dataset Publication

As is obvious from the previous sections, the number of datasets recently published in the area of scientific peer review analysis is already considerable, although their re-use is impeded by the non-transparent way in which the data was collected. Therefore, we advocate for the adherence to data creation and publication guidelines such that their reuse is facilitated. Making sure that such datasets are findable, that they are well documented, and additionally, that machine-readable meta-data about them is available are ways to contribute towards their reuse. The current list of available datasets published following the above-mentioned Dagstuhl seminar[8] is a first step toward achieving this, however incomplete or lacking machine-interpretable data and meta-data. Some of the recommendations we make in the following subsections are, in fact, aligning with the FAIR principles for data sharing[9].

Furthermore, whenever possible, we recommend researchers to describe their datasets through appropriate metadata schema[10].

4.1 Dataset Reuse

As discussed in the previous section, there are many overlaps between existing datasets. Although existing datasets may not fit all research questions, there are several datasets targeting the same task that overlap significantly (Shen et al., 2022; Zeng et al., 2023; Sun et al., 2024b). Therefore, an obvious recommendation is to investigate existing datasets for their appropriateness to the task at hand. Ideally, such datasets are visible to search portals like DataCite[11] or OpenAire[12]. Overview studies like our work or Kuznetsov et al. (2024) are for now an alternative. If necessary these datasets should be extended instead of creating a new dataset from similar underlying sources.

[8]https://github.com/OAfzal/nlp-for-peer-review

[9]FAIR: Findable, Accessible, Interoperable, Re-usable https://www.gofair.foundation/interpretation

[10]E.g. schema.org/docs/data-and-datasets.html

[11]https://datacite.org/

[12]https://www.openaire.eu/

	ICLR								
Dataset	**2013-2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**Ann.**
Kang et al. (2018)		✓							
Wang and Wan (2018)		✓	✓						✓
Hua et al. (2019)		✓	✓						✓
Li et al. (2019)	✓	✓							
Cheng et al. (2020)	✓(-2015)	✓	✓	✓	✓				✓
Chakraborty et al. (2020)		✓	✓	✓					✓
Wang et al. (2020)		✓	✓	✓	✓				✓
Singh et al. (2021)		✓	✓	✓	✓				✓
Choudhary et al. (2021)			✓						✓
Bharti et al. (2021)		✓	✓	✓	✓				
Fromm et al. (2021)				✓	✓				✓
Ghosal et al. (2022a)			✓						✓
Kennard et al. (2022)				✓	✓				✓
Shen et al. (2022)			✓	✓	✓	✓			✓
Yuan et al. (2022)		✓	✓	✓	✓				✓
Fernandes and Vaz-de Melo (2022)				✓	✓	✓			
Wu et al. (2022)		✓	✓	✓	✓	✓			✓
Verma et al. (2022)			✓						✓
Zhang et al. (2022)		✓	✓	✓	✓	✓	✓		
Ghosal et al. (2022b)			✓						✓
Bharti et al. (2023)	?	?	?	?	?	?			✓
Kumar et al. (2023)		✓	✓	✓	✓				✓
Szumega et al. (2023)	?	?	✓	✓	✓	✓			
Wang et al. (2023)		✓	✓	✓	✓	✓	✓		
Fernandes and Vaz-de Melo (2023)			✓	✓		✓			
Li et al. (2023)			✓	✓	✓	✓	✓		✓
Zeng et al. (2023)	?	?	?	?	?	?	?		✓
Huang et al. (2023)							✓		
Purkayastha et al. (2023)			✓	✓	✓				✓
Verma et al. (2023)		✓	✓	✓					✓
D'Arcy et al. (2023)	?	?	?	?	?	?	?		✓
Liu et al. (2023a)							✓		
Jourdan et al. (2024)	✓	✓	✓	✓	✓	✓	✓	?	✓
Sun et al. (2024b)			✓	✓	✓	✓	✓		✓
Zhou et al. (2024)								✓	✓
Sum	3	16	24	21	18	10	7	1	25

Table 9: ICLR data used as source for Peer Reviewing Tasks, per year, marking which datasets were annotated. A question mark indicates that the data was not disclosed.

4.2 Preprocessing and Annotation

When data collected from the open review portals is processed or annotated, details about the processing should be made available. This includes annotation guidelines, inclusion or exclusion criteria for the data points in the final dataset, software. This type of documentation is relevant to researchers who need to understand how data was obtained in order to decide its appropriateness to their task. We found little evidence of preprocessing and annotation documentation. This issue was confirmed, for example by Bharti et al. (2021), who mentioned the lack of documented preprocessing on the datasets when reporting on duplicates and empty reviews in the ICLR 2017 data.

4.3 Availability and Documentation

A key factor for dataset reuse is that it is actually available and the data is provided in an easy to process form, such that integrating it into other processing pipelines needs little overhead.

Many of the datasets that we refer to are available for download, for the remaining, however, either only data acquisition scripts (Geldsetzer et al., 2023) or insufficient selection criteria and indica-

tion of review platform source are provided (Matsui et al., 2021; Lin et al., 2023b; Liu and Shah, 2023; Sun et al., 2024b). In some cases we have found outdated links (Li et al., 2023).

Besides data availability, another issue is the missing documentation of the datasets contents. Several of the published datasets state that they investigated publications from the OpenReview platform, but do not elaborate on which conference and years they use (Bharti et al., 2023; Zeng et al., 2023; D'Arcy et al., 2023; Szumega et al., 2023; Jourdan et al., 2024), making that research difficult to reproduce, if at all possible.

Therefore, we recommend clearly stating the selection criteria for the data (e.g. conferences, years, platforms), whether the data is built on previous datasets (e.g., Guo et al. (2023)), and report descriptive statistics on the released collection, not only for the full dataset, but also for slices of the data (see Sun et al. (2024b)).

4.4 Canonical Splits

The datasets we reported on can be, and are, used both for analysis and for training machine learning models. To make the results of classification or generation tasks comparable, in the classical ML tradition, such datasets should have predefined canonical splits for training, validation, and testing. PeerRead (Kang et al., 2018), ORSUM (Zeng et al., 2023), PolitePeer (Bharti et al., 2023), SubstanReview (Yuan et al., 2022) do contain such predefined splits, but others do not (e.g. Bharti et al. (2021); Kumar et al. (2023); Purkayastha et al. (2023)).

Thus, we recommend to define canonical data splits to make model performance comparable.

4.5 Licensing

As already stated in, for example, (Kuznetsov et al., 2022; Dycke et al., 2022, 2023a) many of the datasets here are not licensed and ethically questionable, as reviewers may not have the option to opt-out of their reviews being published. Furthermore, many of the datasets are published without a clear license, although the original data is licensed. An example of this is the OpenReview platform, which is licensed as CC-BY 4.0[13], but the datasets built on it rarely mention or re-apply this license.

Thus, we advise using clear licenses for published data, and pro-active data collection whenever possible, as proposed by (Dycke et al., 2022).

[13] https://openreview.net/legal/terms

5 Discussion

The recent Dagstuhl Proceedings, *What Can Natural Language Processing Do for Peer Review*, suggest that it is mostly up to researchers in the NLP domain to solve many tasks until AI-assisted Peer Reviewing or Automated (Peer) Reviewing becomes possible. While this is true for most of the tasks in Section 2, there are challenges that, we believe, will profit from solutions that don't rely solely on Large Language Models or Text Analysis.

In our work on this paper, we note that research in peer review (Liu and Shah, 2023; D'Arcy et al., 2024; Ghosal et al., 2019; Fytas et al., 2021; Sun et al., 2024a) aim to automatically generate or predict peer reviews, leaving out the use of domain knowledge to measure the impact, novelty or compare the results to previous research results that use same datasets. One can argue that the knowledge is already encoded in LLMs, we comment that its use is not tractable. To our knowledge, ReviewRobot (Wang et al., 2020) is the only tool that uses Controlled Knowledge-Driven-Generation, to generate reviews using domain knowledge by accessing multiple knowledge graphs.The quality of the results is, however, heavily dependent on the quality of the information extraction processes that provide the content of the knowledge graphs.

In the IR community, information extraction is a long-standing research task, the latest efforts for fine-grained entity extraction being visible in the organization of the RUFEERS task at TREC[14], and the State-of-the-Art task at CLEF[15]. We hope that research in information extraction may contribute to automated population of knowledge graphs, noting that the most extensive knowledge graph (Open Research Knowledge Graph) (Kabongo et al., 2024), which supports comparing scientific publications by various parameters (including datasets), is currently curated by hand.

In conclusion, we believe connecting the research on LLMs with Information Extraction as pursued by the IR community or the Semantic Web community, can help extract the contributions of papers and compare their results to previous ones and effectively support reviewers in their work.

6 Conclusion

In this work, we gave an overview of 53 different datasets for scientific peer reviewing and the tasks

[14] https://tac.nist.gov/2024/RUFEERS/

[15] https://simpletext-project.com/2024/en/

these datasets have been created for. We found several issues with these datasets, such as poor documentation, overlapping data, unclear licensing, and a lack of canonical splits. For this reasons we propose several guidelines for publishing new datasets in this domain, guidelines that, in fact, apply to publishing datasets in general. We also see a need for the creation of a taxonomy of tasks in AI-assisted or automated peer reviewing, and the need for a well-documented dataset that consists of peer reviews from various domains, which can be used for different subtasks.

Currently, the analysis of peer review texts and tool or model development has been done by researchers in the NLP domain. We hope that researchers in other Computer Science fields will assist with their expertise. The Semantic Web community or the Information Retrieval community, are currently only slightly involved in this process. Therefore, we believe that it is necessary to get these communities more involved in this research. This can be done via future shared-tasks, and, although we see the irony, new datasets and tasks that are tailored to make use of the expertise in these domains.

Acknowledgments

While this paper critiques the numerous datasets available, we recognize that creating a dataset is an extremely challenging task. Therefore, as a community, we should be thankful and acknowledge the efforts of everyone who creates and shares datasets with the community.

References

Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. 2023. PolitePEER: does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation*.

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*, pages 421–435. Springer International Publishing.

Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, pages 207–216. Association for Computing Machinery.

Haotian Chen, Han Zhang, Houjing Guo, Shuchang Yi, Bingsheng Chen, and Xiangdong Zhou. 2023. SALAS: Supervised aspect learning improves abstractive multi-document summarization through aspect information loss. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 55–70. Springer Nature Switzerland.

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

Kwangsu Cho. 2008. Machine classification of peer comments in physics: 1st international conference on educational data mining, EDM 2008. *Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings*, pages 192–196.

Gautam Choudhary, Natwar Modani, and Nitish Maurya. 2021. ReAct: A review comment dataset for actionability (and more). In *Web Information Systems Engineering – WISE 2021*, pages 336–343. Springer International Publishing.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. ARIES: A corpus of scientific paper edits made in response to peer reviews.

Iddo Drori and Dov Te'eni. 2024. Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109.

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. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023a. 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. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023b. Overview of PragTag-2023: Low-resource multi-domain pragmatic tagging of peer reviews. In *Proceedings of the 10th Workshop on Argument Mining*, pages 187–196. Association for Computational Linguistics.

Gustavo Lúcius Fernandes and Pedro O. S. Vaz-de Melo. 2022. Between acceptance and rejection: challenges for an automatic peer review process. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, pages 1–12. Association for Computing Machinery.

Gustavo Lúcius Fernandes and Pedro O. S. Vaz-de Melo. 2023. Enhancing the examination of obstacles in an automated peer review system. *International Journal on Digital Libraries*.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021. Argument mining driven analysis of peer-reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4758–4766. Number: 6.

Panagiotis Fytas, Georgios Rizos, and Lucia Specia. 2021. What makes a scientific paper be accepted for publication?

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1274–1290. Association for Computational Linguistics.

Pascal Geldsetzer, Markus Heemann, Pauli Tikka, Grace Wang, Marika Mae Cusick, Ali Lenjani, and Nandita Krishnan. 2023. Prevalence of short peer reviews in 3 leading general medical journals. *JAMA Network Open*, 6(12):e2347607.

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022a. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, 17(1):e0259238. Publisher: Public Library of Science.

Tirthankar Ghosal, Kamal Kaushik Varanasi, and Valia Kordoni. 2022b. HedgePeer: a dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, pages 1–5. Association for Computing Machinery.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130. Association for Computational Linguistics.

Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. 2023. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10198–10216. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137. Association for Computational Linguistics.

Junjie Huang, Win-bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. 2023. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427.

Steven Jecmen, Nihar B. Shah, Fei Fang, and Leman Akoglu. 2024. On the detection of reviewer-author collusion rings from paper bidding.

Steven Jecmen, Minji Yoon, Vincent Conitzer, Nihar B. Shah, and Fei Fang. 2023. A dataset on malicious paper bidding in peer review.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arxivedits: Understanding the human revision process in scientific writing.

Leane Jourdan, Florian Boudin, Nicolas Hernandez, and Richard Dufour. 2024. CASIMIR: A corpus of scientific articles enhanced with multiple author-integrated revisions.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2024. ORKG-leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. *International Journal on Digital Libraries*, 25(1):41–54.

Amir R. Kachooei and Mohammad H. Ebrahimzadeh. 2022. Editorial: What is peer review? *Archives of Bone and Joint Surgery*, 10(1):1–2.

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. Association for Computational Linguistics.

Neha Nayak 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. Association for Computational Linguistics.

Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16693–16704. 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. 2024. What can natural language processing do for peer review?

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112. Association for Computational Linguistics.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. 2019. A neural citation count prediction model based on peer review text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4914–4924. Association for Computational Linguistics.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023a. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023b. MOPRD: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206.

Meng-Huan Liu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023a. ContributionSum: Generating disentangled contributions for scientific papers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5351–5355. ACM.

Ryan Liu, Steven Jecmen, Vincent Conitzer, Fei Fang, and Nihar B. Shah. 2023b. Testing for reviewer anchoring in peer review: A randomized controlled trial.

Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? an exploratory study on using large language models for paper reviewing.

Akira Matsui, Emily Chen, Yunwen Wang, and Emilio Ferrara. 2021. The impact of peer review on the contribution potential of scientific papers. *PeerJ*, 9:e11999.

Dominik Pfütze, Eva Ritz, Julius Janda, and Roman Rietsche. 2022. A corpus for suggestion mining of german peer feedback. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5539–5547. European Language Resources Association.

Barbara Plank and Reinard van Dalen. 2019. CiteTracked: A Longitudinal Dataset of Peer Reviews and Citations. In *BIRNDL@SIGIR*.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495. Association for Computational Linguistics.

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

Michael Ridenour, Ameeta Agrawal, and Olubusayo Olabisi. 2022. Assessing inter-metric correlation for multi-document summarization evaluation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 428–438. Association for Computational Linguistics.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699. Association for Computational Linguistics.

Anna Severin, Michaela Strinzel, Matthias Egger, Tiago Barros, Alexander Sokolov, Julia Vilstrup Mouatt, and Stefan Müller. 2023. Relationship between journal impact factor and the thoroughness and helpfulness of peer reviews. *PLoS biology*, 21(8):e3002238.

Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike von Luxburg. 2018. Design and analysis of the nips 2016 review process.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535. Association for Computational Linguistics.

Shruti Singh, Mayank Singh, and Pawan Goyal. 2021. COMPARE: A taxonomy and dataset of comparison discussions in peer reviews.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2020. Catch me if i can: Detecting strategic behaviour in peer assessment.

Ivan Stelmakh, John Wieting, Graham Neubig, and Nihar B. Shah. 2023. A gold standard dataset for the reviewer assignment problem.

Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024a. ReviewFlow: Intelligent scaffolding to support academic peer reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 120–137. Association for Computing Machinery.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024b. MetaWriter: Exploring the potential and perils of AI writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction*, 8:94:1–94:32.

Jaroslaw Szumega, Lamine Bougueroua, Blerina Gkotse, Pierre Jouvelot, and Federico Ravotti. 2023. The open review-based (ORB) dataset: Towards automatic assessment of scientific papers and experiment proposals in high-energy physics.

Rajeev Verma, Tirthankar Ghosal, Saprativa Bhattacharjee, Asif Ekbal, and Pushpak Bhattacharyya. 2023. ReviVal: Towards automatically evaluating the informativeness of peer reviews. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, pages 95–103. Association for Computing Machinery.

Rajeev Verma, Rajarshi Roychoudhury, and Tirthankar Ghosal. 2022. The lack of theory is painful: Modeling harshness in peer review comments. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 925–935. Association for Computational Linguistics.

Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. 2023. What have we learned from OpenReview? *World Wide Web*, 26(2):683–708.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 175–184. Association for Computing Machinery.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397. Association for Computational Linguistics.

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*, CIKM '22, pages 2189–2198. Association for Computing Machinery.

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

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Scientific opinion summarization: Meta-review generation with checklist-guided iterative introspection.

Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach.

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. ELRA and ICCL.