Generative approach to pragmatics conformation a case study of conference review analysis

Julien Perez Idir Benouaret
Laboratoire de Recherche de l'EPITA, Kremlin Bicètre, France
firstname.lastname@epita.fr

RÉSUMÉ

La relecture en double aveugle est centrale dans les conférences scientifiques, mais des biais persistent. OpenReview a introduit plus de transparence en rendant publics les articles, les évaluations et les décisions. Ce travail explore l'utilisation des grands modèles de langage (LLMs) pour assister différentes étapes du processus de relecture : production de méta-revues, détection de biais et de subjectivité dans les évaluations. L'étude s'appuie sur les données ICLR de 2017 à 2022 et inclut des analyses quantitatives et des évaluations humaines à l'aveugle. Les résultats visent à encourager une relecture scientifique plus efficace et équitable.

ABSTRACT

Generative approach to pragmatics conformation a case study of conference review analysis

Double-blind peer review is central to scientific conferences, yet biases persist. OpenReview has introduced greater transparency by making papers, reviews, and decisions publicly available. This work explores the use of large language models (LLMs) to support various stages of the peer review process : generating meta-reviews, detecting bias and subjectivity in evaluations. The study draws on ICLR data from 2017 to 2022 and includes both quantitative analyses and blind human assessments. The results aim to promote a more effective and fair scientific review process.

MOTS-CLÉS : review pair à pair, bias, grand modèles de langue, OpenReview, ICLR, évaluation par les pairs, revue ouverte.

KEYWORDS: Peer review, Bias, Large Language Models, OpenReview, ICLR, Open review.

1 Introduction

The progress of contemporary science is significantly dependent on the peer review mechanism, enabling authors to publish and disseminate their research findings. Therefore, it is crucial that the peer review process maintains fairness and impartiality, especially for emerging researchers whose career paths are significantly influenced by paper acceptance. Regrettably, the peer review process has recently come under scrutiny. Due to the overwhelming volume of submissions, many platforms have faced an unprecedented demand for reviewers, resulting in numerous complaints from authors regarding unfair or substandard reviews. As an example, the NeurIPS experiments [Cortes and Lawrence, 2021] [1] assigned a different set of reviewers to the same submissions, and found the reviewer ratings are often significantly different. Then, [McGillivray and De Ranieri, 2018] analyzed

[1] Both in 2014 [Cortes and Lawrence, 2021] and 2021, see https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/.

128,454 articles in Nature-branded journals and found that authors from less prestigious academic institutions are more likely to choose double-blind review as opposed to single-blind review.

In this context, LLMs have attracted attention of various domains to analyze the intrinsic bias associated to corpus [Lin et al., 2024, Ganguli et al., 2023] and investigated their capabilities to make decisions of diverse nature based on natural language [Liu et al., 2023, Carta et al., 2023]. In this paper, we question whether one could use the zero-shot capabilities of LLMs to assist the peer review process in terms of acceptance and meta-reviewing capabilities in addition to detection of problematic reviews. For this purpose, we utilize a collection of submissions of the International Conference on Learning Representations (ICLR) and propose a series of quantitative and qualitative evaluations.

Contributions. Our contributions are summarized as follows (1) We evaluate the capability of several open-source LLMs for submission acceptance based on original reviews, meta-review and automatically generated meta-reviews. (2) We evaluate an automatic and comprehensive protocol to assess objectivity and bias in a peer review process. (3) We evaluate the zero-shot capabilities of LLMs to detect problematic reviews.

2 Related Work

The peer review dynamics have recently attracted attentions in the NLP community. As a first instance, dataset of papers have been recently created from ACL, NeurIPS, and ICLR venues Kang et al. [2018], Plank and van Dalen [2019]. Another dataset Gao et al. [2019] has focused on investigating the effect of rebuttals in NLP conferences. Then, many recent research has been proposed to investigate the bias present in the review process [Stelmakh et al., 2019, Manzoor and Shah, 2021], argument mining [Hua et al., 2019], automatic review generation [Yuan et al., 2021], improved review process [Rogers and Augenstein, 2020, Jecmen et al., 2020], and review explanation [Wang et al., 2020]. In this context, evaluating the capability of automatic decision of LLMs in this review process hasn't been investigated so far.

Pragmatics and discourse analysis hold pivotal positions within linguistics [Allwood, 1981]. Specifically, pragmatics delves into the contextual nuances of meaning that are often overlooked in the formulation of content or logical structure [Horn and Ward, 2004]. A recent work [Hu et al., 2023] compares human and model comprehension, revealing language models' struggles with humor, irony, and conversational maxims but seems to well capture discourse structure and basic common-sense. Notably, this paper constitutes a novel exploration of pragmatic aspects within conference reviews.

As a second contribution, we analyse the capabilities of large language model to detect such flaws during the scientific review process in addition to acting as an impartial decision maker based on original reviews, meta-reviews and generated meta-reviews. As far as our knowledge goes, none of the current related works has investigated such capabilities through the lens of large language models.

3 ICLR papers corpus

In this study, we utilize a subset of an ICLR corpus of papers. To compile this dataset, we use the `OpenReview` API[2] to gather data from $10,289$ submissions to the International Conference on Learning Representations (ICLR) spanning the years 2017 to 2020. The gathered data encompasses (i) submissions and their associated metadata ; (ii) double-blinded review and rebuttal data ; and (iii) decisions made by area chairs and program chairs. In this work, no information about authorship and associated institutions are considered. The review data is comprised of ratings, confidence levels, and textual reviews and rebuttals. The review and rebuttal data includes each reviewer's comments on the submission, a numerical score, a confidence rating, and the authors' rebuttal. Lastly, we binarize the paper decision by categorizing diverse acceptance designations, i.e spotlight, poster, short talk, into two groups : *Accept* and *Reject*. Upon paper assignment, a minimum of three reviewers independently evaluate a paper. Following the rebuttal phase, reviewers can access the authors' responses and other peer reviews, enabling them to revise their reviews accordingly. The program chairs then compile a meta-review for each paper, making the final accept/reject decision based on the three anonymous reviews. Each official review primarily consists of a review score, a reviewer confidence level, respectively from 1 to 10 and 1 to 5, and detailed review comments. The final dataset we are sampling from is composed with 5,527 submissions and 16,853 official reviews from ICLR 2017-2020 venues on the OpenReview platform. Desk-rejected and withdraw submissions have been pruned.

4 Large Language Models for Reviews

Leveraging this ICLR corpus, we aim at answering the following questions : **(RQ1)** How can LLMs be used to generate accurate acceptance decisions directly from the list of reviews, meta-review and generated meta-review from the set of original reviews ? **(RQ2)** How do LLMs compare to existing dedicated models of sentence-level polarity, bias and subjectivity detection for conference review evaluation ? **(RQ3)** Can LLMs be used to identify reviews that require special attention in regard of subjectivity and bias ? Importantly, our approach leverages the capabilities of LLMs to analyse discourse structures and contents. As a consequence, we assume that the model requires only limited knowledge about the actual subject of the paper to establish the feedbacks and decisions while remaining informative, objective and impartial. For each experiment, the prompt is given in appendix. For experiments involving annotators, the participants have been asked to rate the model's decision on each review with a discrete value of $\{-1, 0, 1\}$ to express their accordance with the model's decision. A score of -1 indicates disagreement, 0 indicates neutrality, and 1 indicates agreement.

4.1 Paper acceptance and Feedbacks

We evaluate LLM's of diverse parameter sizes to make acceptance decisions based on the original set of reviews and meta-review. Furthermore, we assess LLM's to provide meta-review which are usable for acceptance decision.

[2] https://api.openreview.net/api/

Method	Models				
	Gemma-7b	Starling-7b	Mistral-7b-instruct	Llama3-8b	Mixtral-7x8b
Review	0.79	0.85	0.63	0.74	.85
MetaReview	0.50	0.45	0.51	0.51	0.48
Gen-MetaReview	**0.8**	**0.93**	**0.76**	**0.8**	**0.83**

TABLE 1 – Acceptance accuracy based on review, metareview and generated metareview for a series of instructed-LLMs.

Acceptance decision We assess the performance of several LLMs in making acceptance decisions based on three different sources of information : the set of original reviews, the meta-review and a generated meta-review produced from the original reviews. Table 1 compares the LLM's decisions with the actual final acceptance decisions. A first noticeable result is the capabilities of the model to predict decision. In this evaluation, we use a balanced dataset to accept and reject paper. Across the models, the generated meta-review method consistently yields the highest accuracy scores, with Starling-7b achieving the highest at 0.93. In contrast, the MetaReview method shows the lowest accuracy across all models, indicating that generating metareviews significantly improves acceptance accuracy compared to relying solely on the original reviews or metareviews. One surprising result was the difficulty of the model to leverage only the meta-review. This phenomenon can be explained by noting that meta-reviews often fail to compile the exhaustive set of feedback provided in individual reviews, rendering the decision-making process intractable. For this reason, we decided to prompt each LLM to generate a meta-review based on the original reviews and then test the model on acceptance. One noticeable result is that the generated meta-reviews significantly improve the acceptance rate. Additionally, we observe that this summarizing process helps the model make better decisions compared to working with the longer context of the original set of reviews.

Meta-review generation : Following on the positive results of acceptance prediction from one generated meta-review, we conduct a first human-study to assess model's ability to summarize reviews. We assess the quality of the generated meta-reviews based on several criteria. In this setting, the models were provided with the set of reviews for a given paper and were tasked with generating a meta-review that effectively captured the main points, arguments, and findings from these reviews. We then asked human evaluators to rate the generated meta-reviews comparing them with the real meta-reviews written by human experts. As an evaluation, we compare the generated meta-review and the actual one on five criteria. For each meta-review, the annotator is asked the determine its preference to the original one. We use the following criterias : Clarity, Accuracy, Comprehensiveness, Fairness, Overall. The details of each criteria is provided in Appendix. One can observe that Mixtral-7x8b performed best across most criteria, with scores of 0.21±0.66 for clarity, 0.28±0.58 for accuracy, 0.14±0.78 for comprehensiveness, and 0.24±0.62 for fairness. Additionally, Mixtral-7x8b and Llama3-8b tied for the best performance in preference rate with scores of 0.10±0.76 and 0.10±0.71, respectively. This result suggests the necessity to investigate with larger models in further experiments to determine if increased model size can lead to even more significant improvements in generating high-quality meta-reviews.

Criterias	Gemma-7b	Starling-7b	Mistral-7b-instruct	Llama3-8b	Mixtral-7x8b
Clarity	0.05±0.39	0.03±0.56	0.14±0.43	0.34±0.48	**0.21±0.66**
Accuracy	0.11±0.31	-0.14±0.57	-0.28±0.58	0.03±0.61	**0.28±0.58**
Comprehensiveness	0.05±0.39	0.00±0.53	0.07±0.36	0.28±0.45	**0.14±0.78**
Fairness	-0.05±0.39	-0.31±0.53	-0.24±0.57	-0.03±0.67	**0.24±0.62**
Preference	-0.11±0.31	-0.10±0.61	-0.21±0.61	0.10±0.71	**0.10±0.76**

TABLE 2 – Human Evaluation of large language models' performance in generating meta-reviews based on clarity, accuracy, comprehensiveness, fairness, and preference rate in comparison to real meta-reviews

4.2 Reviews

In this second part, we assess the ability of large language models (LLMs) to analyze individual reviews. We aim to measure their capacity to identify and evaluate several aspects of a review, comparing their performance to a series of models fine-tuned for specific tasks. The aspects we focus on are the objectivity, polarity, and biases of individual sentences within a review.

To compare the LLM's performance with these fine-tuned models, we conducted an automatic assessment of accuracy. We computed the accuracy scores between each fine-tuned model and the zero-shot predictions of the LLM on a common set of individual reviews. This comparison allowed us to determine how well the LLM could identify and analyze relevant aspects of a review without any task-specific fine-tuning, using specialized models explicitly trained for these tasks as benchmarks.

For subjectivity, we use the state of the art model deBERTaV3 [He et al., 2021] which is a pretrained masked language model fine-tuned for this comprehension task. For evaluating bias, we use Roberta [Liu et al., 2019] pretrained weights, with a classification head fine-tuned to classify text into two categories (neutral, biased). For this second model, the dataset used to fine-tune the model is extracted from an annotated version of the English wikipedia corpus [Pryzant et al., 2019]. Finally, for polarity, The reference model is a fine-tuned version of BERT [Devlin et al., 2019] on the amazon polarity dataset [Zhang et al., 2015]. Table 3 detailed the results obtained on each LLM's. The results suggest that Llama3-8b stands out as a top performer across all criterias. The tie between Llama3-8b and Mixtral-7x8b in polarity detection further underscores the strengths of these models in understanding the emotional tone of text.

Models	**Evaluation Criteria**		
	Objectivity	**Biases**	**Polarity**
Gemma-7B	0.59	0.50	0.81
Mistral-7B-Instruct	0.66	0.67	0.75
Llama3-8b	**0.75**	**0.74**	**0.83**
Mixtral-7x8b	0.68	0.71	**0.83**

TABLE 3 – Evaluation results for language models based on objectivity, biases, and polarity metrics.

Models	Agreement
Gemma-7b	-0.16±0.49
Starling-7b	0.24±0.57
Mistral-7b-instruct	0.07±0.64
Llama3-8b	0.03±0.72
Mixtral-7x8b	**0.34±0.60**

TABLE 4 – Evaluation for each Language Model on agreement to Human for detection of problematic reviews.

4.2.1 Problematic Reviews

Finally, we evaluate the LLM's ability to identify reviews that require special attention, are of poor quality, or are misleading. Ideally, we aim to establish a consistent and objective method for flagging such reviews for further attention or filtering. To evaluate the LLM's performance in this task, we compared its results to human judgment for a subset of reviews. We defined a zero-shot generation task using a prompt in which each LLM was asked to identify any problems in the discourse and argument structure of the reviews. This subset was curated to include reviews exhibiting various forms of bias, such as confirmation bias, selection bias, or stereotyping, as well as reviews with high levels of subjectivity that could potentially mislead the evaluation process. As detailed in Table 4, Mixtral-7x8b stands out with a score of 0.34±0.60, indicating a reasonable level of agreement with human assessors. This suggests that Mixtral-7x8b is effective in identifying problematic reviews, aligning closely with human judgment. In contrast, Gemma-7b performed poorly with a score of -0.16±0.49, indicating a lack of agreement with human evaluators. The other models, Starling-7b, Mistral-7b-instruct, and Llama3-8b, showed varying degrees of agreement, but none matched the level of Mixtral-7x8b. These findings highlight the need to further explore the capabilities of even larger language models on this task.

5 Limitations and Ethical concerns

Despite the promising results of our study on the evaluation of large language models (LLMs) for writing meta-reviews, deciding on paper acceptances, and detecting problematic reviews, several limitations should be acknowledged. Our evaluation is based on a specific dataset of academic papers and reviews within the research of machine learning, which may limit the generalizability of our findings to other disciplines with distinct writing styles and review standards. The quality and representativeness of the training data significantly influence the model's outputs, potentially leading to biases in model performance. Evaluating the quality of academic reviews involves judgments that LLMs may not fully capture, potentially overemphasizing certain aspects of reviews while underestimating others. Then, identifying problematic reviews remains challenging due to the subjective nature of such determinations, and LLMs might not consistently identify all problematic elements accurately. The use of LLMs also raises ethical concerns about transparency, accountability, and fairness, necessitating responsible and ethical implementation. Furthermore, LLMs require substantial computational resources, which may limit their accessibility and scalability. Future research should aim to address these limitations by expanding the diversity of datasets, improving model robustness.

6 Prompts

In this section, we will detail the prompts used in our paper. Each prompt follows the so-called RISEN, Role/Instruction/Steps/End-goal/Narrow, format that we found working consistently thoughout our experiments. Role : Defines the specific function of the reviewer, meta-reviewer, or program chair ; Instruction : Provides clear directives on the task to be accomplished ; Steps : Outlines the step-by-step process to follow for completing the task ; End Goal : Specifies the objective to be achieved by the end of the task ; Narrow : Sets constraints to ensure focus and relevance in the task execution.

> Role : you are a meta-reviewer for machine learning conference.
> Instruction : Your task is to provide an decision based on the given reviews which are related to the paper title.
> Steps :
> 1. Analyze the provided reviews.
> 2. Identifying common praises and criticisms.
>
> End-Goal : Evaluate the paper in light of the reviews and your own analysis, discussing its strengths and weaknesses.
> Narrow : You must write "Decision : Accept" or "Decision : Reject" only as response.
> The abstract of the paper is as follows : abstract
> The reviews are as follows : reviews

FIGURE 1 – Prompt scheme to produce the acceptance decision based on a collection of original reviews for a given submission.

> Role : you act as a program chair for machine learning conference.
> Instruction : Your task is to provide a decision based on the given meta-review which are related to the paper title.
> Steps : 1. Analyze the provided meta-review. 2. Identifying common praises and criticisms.
> End-Goal : Evaluate the paper in light of the meta-review and your own analysis, discussing its strengths and weaknesses.
> Narrow : As a response, You must write "Decision : Accept" or "Decision : Reject" as answer.
> The abstract of the paper is as follows : {abstract}
> The meta-review is as follows : {chairs}

FIGURE 2 – Prompt to produce the acceptance decision based on a meta-reviews for a given submission.

7 Human Annotation

To ensure the reliability and validity of our evaluation, we employed human annotators to assess the performance of the large language models. Each human annotator was given a series of 10 papers to review, which were randomly selected from the dataset. At the time of submission, we collected annotations from two experienced reviewers who were familiar with the conference review process.

Role : you are a meta-reviewer for machine learning conference.
Instruction : Your task is to provide an analysis of the given reviews which are related to the paper abstract.
Steps :
1. Based on the reviews and abstract, summarize the main findings of the paper, highlighting the authors' approach and the key outcomes.
2. Analyze the provided reviews, identifying common praises and criticisms.
3. Based on your analysis, provide suggestions for improvement.
4. Focus on addressing the concerns raised in the reviews.
End-Goal :
1. Evaluate the paper in light of the reviews and your own analysis, discussing its strengths and weaknesses.
2. Your input should help in understanding the paper's contribution to the field and potential areas for improvement.
Narrow :
1. You must write in plain English.
2. Decide, based on your analysis, if the paper is ready for publication or not.
3. You only act as the meta-review, you are not allow to act as the author of the paper.

FIGURE 3 – Prompt scheme to produce a meta-review based on a given paper abstract and the corresponding individual reviews.

Role : you are analysing the reviews of a machine learning conference.
Instruction : Your task is to provide a sentence-wise analysis of the given review in terms of subjectivity, polarity, biais and topics.
Steps :
1. For each sentence of the given review
2. Extract polarity for each sentence
3. Extract subjectivity for each sentence
4. Extract biais for each sentence
5. Extract topics for each sentence
End-Goal : produce a JSON formatted output with the following structure : [{ "sentence" : "The paper is well-written and the results are convincing.", "polarity" : "positive", "subjectivity" : "objective", "bias" : "neutral", "topics" : ["writing", "results"] },]

FIGURE 4 – Prompt used to produce the evaluation of bias, subjectivity and polarity at a sentence basis.

To maintain objectivity, the model identity was blinded during the evaluation, meaning that the annotators were unaware of which reviews were generated by humans and which were generated by the models. This blinding process helped to minimize any potential biases and ensured that the annotators' assessments were based solely on the content and quality of the reviews. The human annotators' evaluations served as a benchmark against which the performance of the large language models was compared, providing a comprehensive understanding of their capabilities and limitations in generating high-quality reviews.

Role : Assistant to verify is a conference paper review is problematic.
Instruction : Analyze the provided conference paper reviews and estimate if they are problematic.
Steps :
1. Read and understand the provided conference paper reviews.
2. Identify any instances of bias, lack of understanding of the subject matter, subjective opinions presented as facts, and any immoral or unethical statements.
3. Provide a detailed and objective analysis of the reviews, citing specific examples where necessary.
4. Summarize the overall estimation of the reviews, indicating whether they are problematic or not.
End-Goal :
To provide a comprehensive and objective analysis of the conference paper reviews, indicating whether they are problematic or not, and to support the analysis with specific examples from the reviews.
Narrow :
1. The analysis should focus solely on the conference paper reviews provided and should not include any external information or sources.
2. The analysis should also be limited to the specific instances of bias, lack of understanding, subjectivity, and immorality, and should not include any other aspects of the reviews.
The title of the paper is as follows : title
The review to check is as follows : reviews

FIGURE 5 – Prompt to produce the evaluation of overall compliance to standard practices of reviewing and detect problematic reviews.

8 Annotation Instructions for Human Evaluators

Participants in this annotation task are provided the following instructions for evaluating models used in the automated assessment of conference reviews and meta-reviews :

Task Overview : Your participation is crucial for evaluating a small set of language models (LLMs) across two tasks. Each participant is contacted individually to maintain independence. The task involves reviewing and filling Excel sheets with numeric evaluations.

Evaluation Criteria : Each evaluation criterion uses the scale : {1, 0, -1}.

- **1** : Unsatisfactory performance
- **0** : Average performance
- **-1** : Satisfactory performance

First Experiment Evaluate whether you agree with the analysis of the review performed by the LLM. Each file contains original reviews and the LLM's analysis. Provide your assessment in the 3rd column.

Second Experiment Evaluate meta-reviews generated by LLMs based on the following criteria :

- **Clarity :** Assess the clarity and ease of understanding of the meta-review compared to the actual meta-review.
- **Accuracy :** Evaluate how accurately the meta-review reflects the content of individual reviews.
- **Comprehensiveness :** Measure how well the meta-review covers all important aspects discussed in the individual reviews.
- **Fairness :** Determine how fairly the meta-review represents the opinions of all reviewers.

— **Overall Impression :** Provide an overall assessment of the meta-review's quality.
These instructions guide the annotation process aimed at assessing the effectiveness of LLMs in enhancing the objectivity and quality of conference review processes. The definition of the evaluation terms given above where provided at the end of this description.

9 Conclusion

This paper evaluates large language models for academic paper reviewing and meta-reviewing. We assess their ability to decide paper acceptance, generate meta-reviews, and identify problematic reviews. Our findings show that LLMs can effectively support and augment the review process, improving the quality and reliability of assessments. Our results aim to foster more research in this direction.

Références

J. Allwood. On the distinctions between semantics and pragmatics. In *Crossing the Boundaries in Linguistics : Studies Presented to Manfred Bierwisch*, pages 177–189. Springer, 1981.

R. Bhardwaj and S. Poria. Red-teaming large language models using chain of utterances for safety-alignment. *ArXiv*, abs/2308.09662, 2023. URL https://api.semanticscholar.org/CorpusID:261030829.

T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. *ArXiv*, abs/2302.02662, 2023. URL https://api.semanticscholar.org/CorpusID:256615643.

C. Cortes and N. D. Lawrence. Inconsistency in conference peer review : Revisiting the 2014 NeurIPS experiment. Available at https://arxiv.org/abs/2109.09774, 2021. URL https://arxiv.org/abs/2109.09774.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

D. Ganguli, A. Askell, N. Schiefer, T. Liao, K. Lukovsiut.e, A. Chen, A. Goldie, A. Mirhoseini, C. Olsson, D. Hernandez, D. Drain, D. Li, E. Tran-Johnson, E. Perez, J. Kernion, J. Kerr, J. Mueller, J. D. Landau, K. Ndousse, K. Nguyen, L. Lovitt, M. Sellitto, N. Elhage, N. Mercado, N. Dassarma, R. Lasenby, R. Larson, S. Ringer, S. Kundu, S. Kadavath, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. B. Brown, C. Olah, J. Clark, S. Bowman, and J. Kaplan. The capacity for moral self-correction in large language models. *ArXiv*, abs/2302.07459, 2023. URL https://api.semanticscholar.org/CorpusID:256868727.

Y. Gao, S. Eger, I. Kuznetsov, I. Gurevych, and Y. Miyao. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1274–1290. Association for Computational Linguistics, 2019.

P. He, J. Gao, and W. Chen. Debertav3 : Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543, 2021. URL https://api.semanticscholar.org/CorpusID:244346093.

L. R. Horn and G. L. Ward. *The handbook of pragmatics*. Wiley Online Library, 2004.

J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. L. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4194–4213. Association for Computational Linguistics, 2023. DOI : 10.18653/v1/2023.acl-long.230. URL https://doi.org/10.18653/v1/2023.acl-long.230.

X. Hua, M. Nikolov, N. Badugu, and L. Wang. Argument mining for understanding peer reviews. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2131–2137. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/n19-1219.

S. Jecmen, H. Zhang, R. Liu, N. B. Shah, V. Conitzer, and F. Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/93fb39474c51b8a82a68413e2a5ae17a-Abstract.html.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts. *ArXiv*, abs/2401.04088, 2024. URL https://api.semanticscholar.org/CorpusID:266844877.

D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz. A dataset of peer reviews (PeerRead) : Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana, 2018. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/N18-1149.

L. Lin, L. Wang, X. Zhao, J. Li, and K.-F. Wong. Indivec : An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. *ArXiv*, abs/2402.00345, 2024. URL https://api.semanticscholar.org/CorpusID:267365176.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta : A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. URL https://api.semanticscholar.org/CorpusID:198953378.

Z. Liu, W. Yao, J. Zhang, L. Xue, S. Heinecke, R. Murthy, Y. Feng, Z. Chen, J. C. Niebles, D. Arpit, R. Xu, P. T. Mùi, H. Wang, C. Xiong, and S. Savarese. Bolaa : Benchmarking and orchestrating llm-augmented autonomous agents. *ArXiv*, abs/2308.05960, 2023. URL https://api.semanticscholar.org/CorpusID:260865960.

E. A. Manzoor and N. B. Shah. Uncovering latent biases in text : Method and application to peer review. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4767–4775. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16608.

B. McGillivray and E. De Ranieri. Uptake and outcome of manuscripts in nature journals by review model and author characteristics. *Research Integrity and Peer Review*, 3(1) :5, 2018.

G. T. T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Riviere, M. Kale, J. C. Love, P. D. Tafti, L. Hussenot, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. H'eliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikula, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, P. G. Sessa, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. B. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy. Gemma : Open models based on gemini research and technology. *ArXiv*, abs/2403.08295, 2024. URL https://api.semanticscholar.org/CorpusID:268379206.

B. Plank and R. van Dalen. Citetracked : A longitudinal dataset of peer reviews and citations. In M. K. Chandrasekaran and P. Mayr, editors, *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019*, volume 2414 of *CEUR Workshop Proceedings*, pages 116–122. CEUR-WS.org, 2019. URL http://ceur-ws.org/Vol-2414/paper12.pdf.

R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang. Automatically neutralizing subjective bias in text. *ArXiv*, abs/1911.09709, 2019. URL https://api.semanticscholar.org/CorpusID:208248333.

A. Rogers and I. Augenstein. What can we do to improve peer review in nlp ? In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics : EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1256–1262. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.findings-emnlp.112.

I. Stelmakh, N. B. Shah, and A. Singh. On testing for biases in peer review. In *ACM EC workshop on Mechanism Design for Social Good*, pages 5287–5297, 2019.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama : Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, and N. F. Rajani. Reviewrobot : Explainable paper review generation based on knowledge synthesis. In B. Davis, Y. Graham, J. D. Kelleher, and Y. Sripada, editors, *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 384–397. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.inlg-1.44/.

W. Yuan, P. Liu, and G. Neubig. Can we automate scientific reviewing ? *CoRR*, abs/2102.00176, 2021. URL https://arxiv.org/abs/2102.00176.

X. Zhang, J. J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Neural Information Processing Systems*, 2015. URL https://api.semanticscholar.org/CorpusID:368182.

10 Large Language Models

Gemma-7b Gemma-7b [Mesnard et al., 2024] is built on a transformer architecture, consisting of 7 billion parameters. This architecture enables the model to perform a wide range of natural language processing tasks by leveraging self-attention mechanisms for efficient text understanding and generation.

Starling-7b Starling-7b [Bhardwaj and Poria, 2023] also utilizes a transformer-based architecture with 7 billion parameters. This model architecture is designed for efficiency and scalability, optimizing the balance between computational resource use and performance. It excels in conversational AI, content generation, and semantic analysis tasks.

Mistral-7b-instruct Mistral-7b-instruct [Jiang et al., 2023] employs a transformer architecture fine-tuned with instructional datasets, featuring 7 billion parameters. This model is specifically optimized for educational and instructional applications, providing clear and accurate instructional output for creating educational content and facilitating interactive learning.

Llama3-8b Llama3-8b [Touvron et al., 2023] is designed using a transformer architecture with 8 billion parameters. This model's architecture supports advanced language understanding and generation capabilities, making it suitable for sophisticated applications such as advanced dialogue systems, creative writing, and complex data interpretation.

Mixtral-7x8b Mixtral-7x8b [Jiang et al., 2024] is a composite model that integrates multiple transformer modules, each with 8 billion parameters. This Mixture-of-Experts architecture allows Mixtral-7x8b to effectively handle diverse and complex tasks by leveraging the strengths of its sub-models, making it particularly effective in multi-faceted problem-solving and cross-domain knowledge application.

11 Evaluation metrics

In this subsection, we detail the criteria used to assess the quality of the meta-reviews generated by our prompts. These evaluation metrics include clarity, accuracy, comprehensiveness, and fairness. Each metric is designed to provide a comprehensive assessment of how well the generated meta-reviews capture the essence of the individual reviews and present them in a coherent, accurate, and balanced manner.

Clarity : Assesses how clear and easy to understand the meta-review is, in comparison to the real meta-review. A well-written meta-review should be coherent, concise, and logically structured, allowing readers to quickly grasp the main points and conclusions.

Accuracy : Evaluates how accurately the meta-review reflects the content of the individual reviews, in comparison to the real meta-review. The meta-review should summarize the key points, arguments, and findings from the reviews without introducing errors, misinterpretations, or distortions.

Comprehensiveness : Measures how well the meta-review covers all important aspects discussed in the individual reviews, in comparison to the real meta-review. A comprehensive meta-review should address the major themes, issues, and questions raised by the reviewers, providing a balanced and thorough overview of their assessments.

Fairness : Determines how fairly the meta-review represents the opinions of all reviewers, in comparison to the real meta-review. The meta-review should give equal weight to each reviewer's perspective, avoiding bias or favoritism towards any particular reviewer or viewpoint.