LLMs as NLP Researchers: Paper (Meta-)Reviewing as a Testbed

Anonymous ACL submission

Abstract

This work is motivated by two key trends. On one hand, large language models (LLMs) have shown remarkable versatility in various generative tasks such as writing, drawing, and question answering, significantly reducing the time required for many routine tasks. On the other hand, researchers, whose work is not only time-consuming but also highly expertise-demanding, face increasing challenges as they have to spend more time reading, writing, and reviewing papers. This raises the question: how can LLMs potentially assist researchers in alleviating their heavy workload?

This study focuses on the topic of `LLMs as NLP Researchers`, particularly examining how effectively LLMs can perform paper (meta-)reviewing. To address this, we constructed the `ReviewCritique` dataset, which includes two types of information: (i) NLP papers (initial submissions rather than camera-ready) with both human-written and LLM-generated reviews, and (ii) each review comes with "deficiency" labels and corresponding explanations for individual segments, annotated by experts. Using `ReviewCritique`, this study explores two threads of research questions: (i) "`LLMs as Reviewers`", how do reviews generated by LLMs compare with those written by humans in terms of quality and distinguishability? (ii) "`LLMs as Metareviewers`", how effectively can LLMs identify potential issues, such as `Deficient` or unprofessional review segments, within individual paper reviews? To our knowledge, this is the first work to provide such a comprehensive analysis.

1 Introduction

Artificial intelligence (AI), particularly through the recent development of large language models (LLMs), has demonstrated remarkable versatility in tasks such as writing, drawing, and question answering (Naveed et al., 2023; Rasool et al., 2024; Kaddour et al., 2023). This has led to significant automation of many time-consuming jobs, potentially replacing more roles with AI. Interestingly, while researchers, the creators of AI/LLMs, benefit from LLMs for simple tasks (Meyer et al., 2023; Altmäe et al., 2023), it still takes years to train a qualified researcher due to the domain-specific and expertise-demanding nature of their work. Researchers now face increasing challenges with more papers to read, to beat, to write, and to review, resulting in longer and more intensive work hours. This raises the question: how promising is the potential for LLMs to work as researchers to alleviate their heavy and somewhat unhealthy workload?

Within the scope of `LLMs as NLP Researchers`, this work focuses on how well LLMs can perform (meta-)reviewing. AI-related conferences and journals are seeing a rapid increase in submissions, making it difficult to recruit enough (meta-)reviewers. Paper reviewers must carefully read submissions and provide comments on the overall story, strengths, weaknesses, writing, etc. The meta-reviewer's responsibility is to ensure the accuracy and constructiveness of the individual review. Therefore, meta-reviewers are expected to be aware of the submission as well as authors' rebuttals, and then assess individual reviews by identifying unreasonable elements and distilling truly constructive comments. There is a latent trend, though debatable and unacknowledged by reviewers, of LLMs participating more frequently in the paper-reviewing process. Therefore, this work explores two research questions: (i) '`LLMs as Reviewers`", how far away or distinguishable are LLM-generated paper reviews from human-written ones? (ii) "`LLMs as Metareviewers`", can LLMs identify `Deficient` review segments by reasoning over the paper submission, other individual reviews, and author rebuttals jointly?

To achieve this, we create the `ReviewCritique` dataset, containing: (i) NLP

papers (original submissions rather than the final camera-ready) with both human-written and LLM-generated reviews, and (ii) each review annotated by NLP experts (most with Ph.D. degrees or area chairing experience) at the sentence level regarding deficiency and professionalism, with explanations. This dataset enables the following analyses.

First, for `LLMs as Reviewers`, we assess the quality of LLM-generated reviews by examining subsections or aspects of the review, such as summary, strengths, weaknesses, writing, etc. We propose a novel metric to measure LLM-generated review diversity across different papers. Our findings indicate that LLMs generate more `Deficient` review segments than human reviewers and often produce paper-unspecific reviews lacking diversity and constructive feedback.

Second, for `LLMs as Metareviewers`, we evaluate LLMs' ability to identify `Deficient` segments in human-written reviews and provide explanations for their judgments. This contrasts with other works treating paper meta-review as a text summarization task given 3+ individual reviews (Li et al., 2023; Shen et al., 2022; Pradhan et al., 2021). We argue that meta-reviewing should be a knowledge-intensive and reasoning-intensive process, with human meta-reviewers being expected to be careful and responsible. We benchmark both closed-source and open-source LLMs on this task, finding that even top-tier LLMs struggle to mimic human experts in assessing individual reviews.

Overall, our contributions are threefold: (i) the `ReviewCritique` dataset with human-written and LLM-generated reviews and fine-grained review deficiency labeling and explanation, serving as a valuable resource for future research on AI-assisted peer review and LLM benchmarking, (ii) the first quantitative comparison of human-written and LLM-generated paper reviews at the sentence level, and (iii) the first analysis of LLMs' potential as both reviewers and meta-reviewers. By highlighting the strengths and limitations of LLMs in scientific peer review, our work paves the way for future works on integrating AI for research.

2 Related Work

Researchers have explored various aspects of AI for reviews. One area of interest is the use of AI to assist in automatically generating peer reviews, such as predicting scores (Li et al., 2020; Zhou et al., 2024; Wang et al., 2020) and writing reviews (Gao et al., 2024; Wang et al., 2020; Yuan et al., 2022; Liu and Shah, 2023) and meta-reviews (Li et al., 2023; Lin et al., 2023). Another line of research focuses on leveraging NLP methods to evaluate the quality of human reviews (Xiong and Litman, 2011; Guo et al., 2023; Kumar et al., 2023; Ghosal et al., 2022b).

To facilitate research on AI for peer review, several datasets have been introduced. PeerRead (Kang et al., 2018), MOPRD (Lin et al., 2023), and NLPeer (Dycke et al., 2023) are datasets containing a large number of peer reviews and their corresponding papers but without expert annotations. Other datasets focus on specific aspects of peer reviews, such as argument (Kennard et al., 2022; Hua et al., 2019; Yuan et al., 2022; Cheng et al., 2020; Ruggeri et al., 2023), politeness (Bharti et al., 2023), uncertainty detection (Ghosal et al., 2022b), contradictions in review pairs (Kumar et al., 2023), and substantiation (Guo et al., 2023). Peer Review Analyze (Ghosal et al., 2022a) annotates reviews across four facets: paper section correspondence, aspect, functionality, and significance. However, these datasets are solely based on reviews and none of them are *highly* expert-demanding. In contrast, `ReviewCritique` is the first dataset to benchmark LLMs' capability as a responsible meta-reviewer.

Recently, researchers have also explored the evaluation of LLMs' deficiency and limitations in automatic paper reviewing tasks (Zhou et al., 2024; Liu and Shah, 2023; Robertson, 2023; Liang et al., 2023). Our work differs from previous works in that we provide a quantitative comparison of human-written and LLM-generated paper reviews at the sentence level. This fine-grained analysis allows us to identify specific areas where LLMs excel or struggle in generating high-quality reviews. We also propose a novel metric to measure LLM-generated review diversity.

3 `ReviewCritique` Curation

In this section, we detail the process of curating `ReviewCritique`, including the criteria for paper selection, the collection of human-written and LLM-generated reviews, the annotation procedure, and the measures taken to ensure data quality.

3.1 Paper Submission & Review Collection

Criteria. We select the papers based on the following criteria: i) Only consider NLP papers; this facilitates the recruitment of sufficient annotators in the NLP domain. ii) Human-written reviews are publicly accessible. iii) Equal distribution of accepted and rejected papers is maintained to inves-

tigate potential review pattern discrepancies based on the final acceptance or rejection of submissions.

From the OpenReview website, we gathered 100 NLP papers (submitted to top-tier AI conferences ICLR and NeurIPS between 2020 and 2023) along with their complete individual reviews (3-5 for each submission), meta-reviews, and author rebuttals. The revision history on OpenReview allowed us to collect the latest paper submissions before the conference deadline, as these versions are the ones on which the reviews are based.

Question: How can we ensure that the collected individual reviews are written by human experts rather than AI? During the subsequent annotation process, we instruct annotators to notify us if they suspect that a review collected here was likely generated by AI; if any doubts arise, we will discard the paper and all its metadata.

Collecting LLM-generated Reviews To directly compare human-written and LLM-generated reviews, we selected a subset of 20 papers from the original 100. The main reason for this selection was the time-consuming nature of subsequent annotation; a size of 20 allowed for an acceptable statistical comparison. This subset of papers also maintains an equal distribution of accepted and rejected papers. We utilized three of the most powerful closed-source LLMs, namely GPT-4 (OpenAI, 2023), Gemini-1.5 (Google, 2023), and Claude Opus (Anthropic, 2024), as these are the models most likely to be used by humans seeking AI assistance in their reviews. Each LLM generated three reviews using prompts that included the ICLR review guidelines, randomly chosen human-written reviews for both accepted and rejected papers, and a generation template in ICLR 2024 format. This prompt can be found in Table 14 (Appendix F).

3.2 Data Annotation

Annotating Criteria for `Deficient`. We, a group of senior NLP researchers with rich Area Chairing experience, define `Deficient` review segments as follows:

- Sentences that contain factual errors or misinterpretations of the submission.
- Sentences lacking constructive feedback.
- Sentences that express overly subjective, emotional, or offensive judgments, such as "*I don't like this work because it is written like by a middle school student.*"
- Sentences that describe the downsides of the submission without supporting evidence, for example, "*This work misses some related work.*"

Question: Why not directly use author rebuttal to infer the `Deficient` review segments? We do not solely rely on author rebuttals for several reasons. First, author rebuttals are not always correct and may overstate contributions or include information not originally presented in the submission. Second, authors sometimes make compromises to satisfy reviewers even when the review is `Deficient`. Third, author rebuttals do not address all `Deficient` details and mainly focus on the "weakness" part, while "`Deficient`" issues can arise in other parts of the reviews.

Annotator Recruiting. Our annotator team consisted of 40 members from the NLP community, all with multiple first-authored publications in top-tier NLP venues and extensive reviewing experience. 16 have Ph.D. degrees, and 11 are university faculty members, 15 have served as area chair (AC, also called meta-reviewer in some venues) before.

Annotation Process. The annotation was conducted on both human-written and LLM-generated reviews, following these steps: i) *Paper Selection*: To ensure high-quality annotations, annotators were allowed to choose papers that aligned with their expertise and interests, ensuring their proficiency in reviewing these papers. ii) *Awareness of Review Scope*: Our assessment focused on reviews written before the rebuttal phase, i.e., reviews based on the original submission. This decision was made to avoid the multi-turn problem and to keep the scope manageable. We did not consider extra experiments conducted during the rebuttal phase, as pre-rebuttal reviews are based on the original submission. Annotators were required to thoroughly read all reviews, meta-reviews, author rebuttals, and the original submission to ensure a comprehensive understanding of the paper and its associated reviews. iii) *Segment-level Annotation*: For detailed analysis, reviews were segmented by sentences, and annotators were asked to label each sentence (a) whether it is `Deficient`, and (b) provide an explanation if it is. This approach allows for the identification of specific sentences that may be `Deficient`, even if the overall review is of high quality. Meta-reviewers are expected to analyze individual reviews sentence by sentence.

Question: Some reviews are generated by LLMs, how did we ensure that annotators were unaware? For the annotation of LLM-generated

	Human-written Review			LLM-generated Review		
	All	**Accepted**	**Rejected**	**All**	**Accepted**	**Rejected**
#Papers	100	50	50	20	10	10
#Reviews	380	195	185	60	30	30
w/ `Deficient` seg.	272	132	140	60	30	30
w/ `Deficient` pct. (%)	71.57	67.69	75.67	100	100	100
#Segments	11,376	6,027	5,349	1,612	798	814
`Deficient`	713	317	396	221	106	115
`Deficient` pct. (%)	6.27	5.26	7.40	13.71	13.28	14.13
#ExplainationTokens	14773	6957	7816	4156	1978	2178

Table 1: Statistics of `ReviewCritique`.

reviews, we employed a separate group of annotators who were not informed that these reviews were LLM-generated. To prevent potential reminders for internet searches, we concealed submission information, such as "Under review as a conference paper at ICLR 2022," in the papers provided to the annotators. We acknowledge that this approach cannot guarantee complete unawareness.

Quality Control. To maintain annotation quality, two annotators independently reviewed each paper's reviews without access to each other's annotations to prevent bias. Disagreements between the two annotators were resolved by a senior expert with area chair (AC) experience, who examined the conflicting annotations and resolved discrepancies by removing or rewriting the explanations for the unconvincing annotations.

Annotation Timeline. Due to the time-consuming nature of high-quality annotation, each annotator was assigned one paper per week, resulting in a six-month data collection period. This ensured thorough and thoughtful annotations. We organized regular meetings to discuss any issues that arose during the annotation process.

3.3 Data Statistics

Table 1 provides the statistics for our `ReviewCritique` dataset. It shows that 71.57% of human-written reviews and 100% of LLM-generated reviews have `Deficient` segments. A comparison of accepted and rejected submissions reveals that rejected papers consistently contain a higher percentage of `Deficient` segments in both human-written and LLM-generated reviews.

3.4 Novelty of `ReviewCritique`

As shown in Table 2, `ReviewCritique` differs from previous works in several key aspects. First, `ReviewCritique` labels review deficiencies at the sentence level, demanding highly experienced annotators. Second, annotators must read the initial submission, meta-reviews, all reviews, and rebuttals before annotating, unlike previous works that require reading reviews and, at most, rebuttals. These differences make `ReviewCritique` the only dataset suitable for benchmarking LLMs as responsible meta-reviewers, offering a comprehensive evaluation of review quality. Additionally, `ReviewCritique` includes expert-annotated LLM-generated reviews, enabling direct comparison between human and LLM-generated reviews at a granular level. These unique features distinguish `ReviewCritique` and open new research opportunities in AI for peer review.

Dataset	PeerRead	PRAnalyze	Subs.PR	DISAPERE	ReviewCrit.
Sentence-level		✓	✓	✓	✓
Initial submission	✓				✓
Highly Expert-demanding					✓
Deficiency Labeling					✓
Human Review	✓	✓	✓	✓	✓
LLM review					✓
Accepted+Rejected	✓	✓			✓

Table 2: Comparison of `ReviewCritique` with PeerRead (Kang et al., 2018), Peer Review Analyze (Ghosal et al., 2022a), Substantiation PeerReview (Guo et al., 2023) and DISAPERE (Kennard et al., 2022).

4 Experiments

We present experimental results and analysis in two threads: `LLMs as Reviewers` (Section 4.1), and `LLMs as Metareviewers` (Section 4.2).

4.1 `LLMs as Reviewers` (i.e., Human-written reviews vs. LLM-generated reviews)

In this section, we compare LLM-generated reviews with human-written reviews: i) by the fine-grained error types if the review segments are annotated `Deficient`, ii) by fine-grained analysis for each component (summary, strengths, weakness,

Error Type	Human (%)	LLM (%)
Human top-3		
Misunderstanding	22.86	10.41
Neglect	19.64	4.52
Inexpert Statement	18.23	5.88
LLM top-3		
Out-of-scope	4.35	31.67
Misunderstanding	22.86	10.41
Superficial Review	2.66	9.95

Table 3: Comparing top-3 error types between human-written and LLM-generated reviews.

writing, and recommendation score), iii) by considering review diversity.

4.1.1 Error type analysis for deficiency

Besides the coarse-grained "`Deficient`" label, our annotation team classify the expert-annotated `Deficient` segments into 23 fine-grained error types (full list and their explanations in Table 9, Appendix D). Table 8 (Appendix D) report the percentage of each error type for both human-written and LLM-generated reviews. Table 3 shows the comparison of the top-3 most frequent error types between human and LLM reviews.

From Table 3, a major reason for `Deficient` reviews from human reviewers is misunderstanding the paper submission and raising unnecessary concerns by neglecting information already stated. This suggests a lack of patience during the reviewing process. Another significant error is making inexpert critiques or statements due to insufficient domain knowledge, potentially from unqualified reviewers being involved due to the increasing number of submissions to AI/NLP conferences and the need to recruit more reviewers.

Compared to humans, LLMs are more likely to suggest out-of-scope experiments or analyses. They make significantly fewer "Inexpert Statement" errors. Based on our observations, this because their reviews are usually paper-unspecific and superficial, avoiding expert-level mistakes. Additionally, LLM-generated reviews do not exhibit errors like "Missing Reference," "Invalid Reference," and "Concurrent Work" since they do not point to specific works or provide references.

4.1.2 Fine-grained review analysis

"Summary" part. The Summary section in LLM-generated reviews exhibits relatively better quality compared to other aspects. Our annotators identified only 1.36% of segments as "Inaccurate Summary" among all LLM `Deficient` segments, which constitutes 0.19% of all LLM-generated segments. In comparison, 5.75% of segments were identified as "Inaccurate Summary" among all `Deficient` segments in human-written reviews, accounting for 0.36% of all human-written review segments. This is nearly twice the percentage found in LLM-generated summaries. Moreover, error types such as "Summary Too Short" and "Copy-pasted Summary", which are present in human reviews, were not observed in LLM-generated reviews, suggesting that LLMs are capable of generating summaries of satisfying quality and avoid directly copying content from the paper.

"Strengths" part. LLMs tend to accept authors' claims in submissions without much critical evaluation. Our analysis reveals that among all segments in the Strengths section of LLM-generated reviews, 53.2% are simply rephrased from the submission, while the remaining segments are mostly inferred from the introduction and abstract, where authors typically highlight their contributions.

To further investigate, we used `ReviewCritique` to compare human-written reviews assessed by annotators and LLM-generated reviews for the same papers. For accepted papers, 34.5% of the Strength segments generated by LLMs were questioned by human experts in their corresponding human-written reviews. For rejected papers, this rose to 51.9%.

These findings suggest that LLMs often accept authors' claims without thorough verification, treating strengths as a text summarization task. In contrast, human reviewers scrutinize the claimed strengths and provide their expert opinions on the validity and significance of the contributions.

"Weaknesses" part. The most dominant type of `Deficient` in LLM reviews is "Out-of-scope", accounting for 31.67% of all `Deficient` segments in LLM-generated reviews (see Table 3). LLMs often highlight weaknesses such as the need for more experiments, lack of generalizability, additional tasks, more analysis, evaluation on languages beyond English, etc. While occasionally relevant, these suggestions often fall outside the paper's scope and shouldn't be considered weaknesses.

Moreover, the suggestions provided by LLMs in the Weaknesses section tend to be paper-unspecific and superficial (e.g, *The paper's focus on pre-trained models might limit its applicability to domains where such models are not available or suit-*

able.), making them applicable to most NLP papers without offering actionable insights to either authors or area chairs. This lack of specificity and depth in the critiques highlights the limitations of LLMs in providing meaningful and constructive feedback on the weaknesses of a paper.

These findings underscore the importance of human expertise in identifying and articulating the most relevant and significant weaknesses of a paper. While LLMs can generate a list of potential limitations, they often struggle to contextualize these weaknesses within the scope and objectives of the paper, leading to `Deficient` segments that may not be helpful to authors or area chairs.

"Writing" part. Our analysis suggests that LLMs may lack the ability to accurately judge the writing quality of a paper submission. In all LLM-generated reviews, LLMs consistently praise the writing of the papers, stating that they are well-written and easy to follow. However, among the papers used for generating LLM reviews, 15% of the papers had both the meta-reviewer and human reviewers agree that the writing was unclear and difficult to follow. Despite this consensus among human experts, the LLMs still provided positive feedback on the writing quality of these papers, failing to accurately assess the writing quality.

"Recommendation Score" part. In addition to generating reviews, we asked LLMs to rate each paper on a scale of 1-10, matching the ICLR and NeurIPS system, for directly comparison with human reviewers. Experiment shows that LLMs tend to give high scores to all submissions, regardless of quality or acceptance status, with averages of 7.43 for accepted and 7.47 for rejected papers. In contrast, human reviewers differentiate more effectively, with averages of 6.41 for accepted and 4.81 for rejected submissions. Thus, LLMs fail to distinguish between accepted and rejected papers, assigning similarly high scores to both.

4.1.3 Review Diversity

Given three LLMs and m papers, we can get a matrix of LLM-generated reviews of size $3 \times m$. We perform quantitative analysis i) horizontally to measure the "intra-LLM review specificity", and ii) vertically as the assessment of "inter-LLM review complementarity".

Intra-LLM Review Specificity. In the real world, we hope the review for each paper is specific to this paper. Then the paper-specific review diversity should discourage two cases: i) one review has too many repeat of certain segment; ii) a review segment appear in too many papers. We get inspiration from the classic TF-IDF to define a new segment-level diversity metric, named **`ITF-IDF`**:

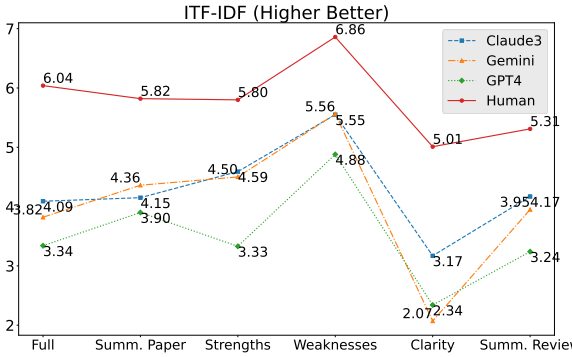

Figure 1: Specificity of reviews: LLM vs. Human.

$$\texttt{ITF-IDF} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\log\left(\frac{n_j}{O_i^j}\right) \times \log\left(\frac{m}{R_i^j}\right)\right), \quad (1)$$

where n_j is the number of segments in review j, O_i^j is the "soft" occurrence of segment s_i^j in review j, R_i^j is the "soft" number of reviews containing segment s_i^j. O_i^j is computed as follows:

$$O_i^j = \sum_{k=1}^{n_j} \mathbb{I}(\text{sim}(s_i^j, s_k^j) \geq t) \cdot \text{sim}(s_i^j, s_k^j), \quad (2)$$

where s_i^j and s_k^j are the i-th and k-th segments in review j, respectively. O_i^j is calculated by summing the similarity scores between segment s_i^j and all other segments s_k^j in the same review j that exceed a predefined similarity threshold t. R_i^j is defined as follows:

$$R_i^j = \sum_{l=1}^{m} \mathbb{I}\left(\max_p \text{sim}(s_i^j, s_p^l) \geq t\right) \cdot \max_p \text{sim}(s_i^j, s_p^l), \quad (3)$$

where s_p^l is any segment in review l. R_i^j is computed by summing the maximum similarity scores between segment s_i^j and segments in each review l that exceed the threshold t. In our experiments, we use SentenceBERT (Reimers and Gurevych, 2019) to calculate the similarity between segments. Implementation details can be found in Appendix A.2.

In summary, `ITF-IDF` measures the specificity of reviews generated by a single LLM across different papers. A lower `ITF-IDF` score means LLM tends to generate repetitive or similar segments across reviews, while a higher score suggests more diverse and unique content in the generated reviews.

Figure 1 shows the Intra-LLM paper-oriented specificity on different review components such as strengths, weaknesses, etc. We set threshold

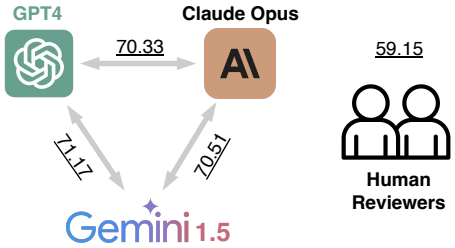

Figure 2: Inter-LLM vs. inter-human review similarities.

t as 0.5 because our initial observation suggests that segments with a similarity higher than this threshold have a similar meaning. We also report the evaluations under different t values in Table 6 (Appendix B). For human-written reviews, we randomly sample one review from each paper and calculate `ITF-IDF`. We repeat this process five times and use the average score.

For `ITF-IDF`, from the full review perspective, human reviews score the highest (6.04), followed by Claude Opus (4.09), Gemini (3.82), and GPT-4 (3.34). The scores are relatively consistent across different sections, but GPT-4 tends to have the lowest scores, suggesting more repetitive segments compared to other LLMs. Human reviews maintain high diversity across all sections. LLMs exhibit a sharp diversity drop in the "Clarity" section. This aligns with our observation in Section 4.1.2 that LLMs praise the writing quality of all papers.

Inter-LLM Review Complementarity. We examine whether different LLMs tend to write complementary reviews for the same paper, which is a pairwise concept. We first compute the BERTScore (Zhang et al., 2020) for each pair of reviews generated by the three LLMs (GPT-4, Claude Opus, and Gemini 1.5) for the same paper. We then average these scores across all papers to obtain an overall measure of Inter-LLM review diversity.

Figure 2 shows the pairwise BERTScores for reviews on the same paper generated by GPT-4, Claude Opus, and Gemini 1.5. It also presents the BERTScores for reviews of the same paper conducted by human reviewers. The BERTScores between different LLM pairs are similar and high, ranging from 70.33 to 71.17. In comparison, the BERTScore between human reviewers is 59.15, which is noticeably lower than the scores between the LLMs. This indicates that human reviewers tend to produce more diverse reviews compared to the LLMs. In addition, this finding implies that the use of multiple LLMs may not necessarily lead to a significant increase in the diversity of perspectives and insights in the review process.

4.2 `LLMs as Metareviewers`

As an area chair, one should assess the quality of individual reviews using their own expertise. This task is highly knowledge-intensive and requires deep understandings of the research domain. Our `ReviewCritique` provides segment-level annotation on if each segment is deficient and why. This section evaluates if prompting popular LLMs (both closed- and open-source) can solve this problem. For closed-source models, we assess GPT4 (OpenAI, 2023), Claude Opus (Anthropic, 2024), and Gemini1.5 (Google, 2023). For open-source models, we evaluate Llama3-8B and -70B (AI@Meta, 2024) and Qwen2-72B (Bai et al., 2023).

To mitigate the impact of prompt-specific performance, we employ two prompting strategies: 1) **`Labeling-All`**: Given everything necessary including a list of indexed review segments, require the LLM to output a list of triples like (id, `Deficient` or not, explanation); 2) **`Select-Deficient`**: Given everything necessary including a list of indexed review segments, require the LLM to output a list of tuples, (id, explanation), when it believes the "id" corresponds to an `Deficient` segment. The detailed prompt templates are in Table 12 and 13 (Appendix F).

To enhance evaluation robustness, we ensemble the results obtained from the two prompting strategies using two methods: i) **Both "No"**: If both prompts classify a segment as `Deficient`, we consider it to be `Deficient`; ii) **Either "No"**: If either of the prompts labels a segment as `Deficient`, we consider it to be `Deficient`.

How well can LLMs identify the `Deficient` segments experts discovered? Metric: we compute the F1 on each paper then average across papers. Table 4 presents the evaluation results.

Closed-source models (GPT-4, Claude Opus, and Gemini 1.5) generally outperform open-source models (Llama3-8B and 70B, Qwen2-72B) in F1 score. Claude Opus achieves the highest F1 scores, with GPT-4 and Gemini 1.5 performing slightly worse. Notably, "recall" scores are consistently higher than precision scores across all LLMs and prompting strategies, suggesting that LLMs tend to incorrectly identify segments as `Deficient`.

Despite the superior performance of the closed-source models, their F1 scores remain relatively low even with different prompt strategies, highlighting the challenges LLMs face in such expertise-

Model	Precision / Recall / F1			
	`Labeling-All`	`Select-Deficient`	Both "No"	Either "No"
GPT-4	14.91 / 34.49 / 18.38	17.18 / 34.59 / 20.30	18.71 / 21.40 / 16.85	14.72 / 47.68 / 20.66
Claude Opus	16.86 / 34.26 / 20.35	17.69 / 26.61 / 18.71	17.14 / 18.70 / 15.78	16.94 / 42.12 / **21.99**
Gemini 1.5	16.58 / 34.13 / 19.76	14.71 / 43.60 / 19.72	17.01 / 27.05 / 18.28	14.46 / 50.37 / 20.34
Llama3-8B	7.73 / 45.95 / 12.22	11.47 / 30.29 / 14.88	11.37 / 21.27 / 12.46	8.19 / 53.61 / 13.35
Llama3-70B	13.63 / 42.49 / 18.19	13.95 / 31.16 / 17.46	16.16 / 23.51 / 16.67	12.46 / 50.02 / 18.43
Qwen2-72B	9.97 / 26.60 / 12.96	11.35 / 34.61 / 14.64	9.07 / 15.13 / 9.62	10.49 / 43.00 / 15.16

Table 4: Performance of LLMs as meta-reviewers on our `ReviewCritique` dataset. The best F1 score among different prompt methods for a single model is underlined. The best F1 score across all models is also **bold**.

Model	ROUGE-1/2/L/BERTScore
GPT-4	17.13 / 2.71 / 14.64 / 55.63
Claude Opus	**20.18 / 3.69 / 17.52 / 57.28**
Gemini 1.5	18.47 / 2.98 / 16.38 / 56.46
Llama3-8B	16.49 / 2.22 / 13.65 / 55.23
Llama3-70B	15.94 / 1.95 / 13.78 / 57.09
Qwen2-72B	17.07 / 3.00 / 14.69 / 56.88

Table 5: Evaluation of LLMs' explanations for correctly identified `Deficient` segments.

intensive tasks and emphasizing the importance of human expertise in the meta-reviewing process.

Can LLMs correctly explain their "`Deficient`" judgment? When LLM's label `Deficient` is correct, we calculate ROUGE (Lin, 2004) and BERTScores between its explanations and our expert's explanations. Table 5 reports evaluation results for the `Select-Deficient` prompt. The full scores for both prompt strategies and their ensembles are in Table 10 and 11 in Appendix E.

The results in Table 5 show that overall scores for all LLMs are relatively low, indicating they can identify some `Deficient` segments but struggle to articulate their reasoning. Among the LLMs, Claude Opus achieves the highest scores across all metrics, suggesting its explanations align best with human annotators. Claude Opus also excels in identifying `Deficient` segments, as shown previously. GPT-4 and Gemini 1.5 show similar performance to Claude Opus. The open-source models, Llama3 (8B and 70B) and Qwen2-72B, generally score lower than the closed-source models.

Which `Deficient` types are challenging for LLMs to identify? To investigate which types of `Deficient` are more challenging for LLMs to detect, we check for each `Deficient` type how many can be successfully identified by LLMs. We focus on three closed-source LLMs: GPT-4, Claude Opus, and Gemini 1.5.

Table 7 (in Appendix C) presents the number and percentage of segments identified in each `Deficient` type by the LLMs. We observe that six types of `Deficient` have a significantly lower percentage compared to the average recall of GPT-4 (47.68%), Claude Opus (42.12%), and Gemini 1.5 (50.37%), suggesting that these types of `Deficient` are particularly difficult for LLMs to detect: `Inaccurate Summary`, `Writing`, `Superficial Review`, `Experiment`, `Contradiction` and `Unstated Statement`

These findings align with our observations in Sections 4.1.2&4.1, where we assessed LLMs as reviewers. For example, LLMs struggle to accurately judge the paper writing quality submission and tend to provide superficial reviews, often failing to offer constructive suggestions on experiments. Moreover, LLMs are more prone to generating contradictory claims in their reviews and making claims that the authors never stated in the submission, indicating a tendency towards hallucination. Additionally, although LLMs can generate paper summaries with fewer errors, they may fail to capture nuanced aspects of the paper, leading to their inability to identify inaccurate summary errors.

5 Conclusion

This work studied the potential of `LLMs as NLP Researchers`, focusing on their roles as reviewers and meta-reviewers. We created `ReviewCritique` , containing both human-written and LLM-generated reviews, with detailed deficiency annotations and explanations. Our analysis reveals that while LLMs can generate reviews, they often produce `Deficient` and paper-unspecific segments, lacking the diversity and constructive feedbacks. Additionally, even state-of-the-art LLMs struggle to assess review deficiencies effectively. These findings highlight the current limitations of LLMs in automating the peer review process.

Limitations

While our work provides valuable insights into the potential of LLMs in the peer review process, there are some limitations to consider. During the evaluation of LLMs, `ReviewCritique` primarily focuses on the textual information from the submissions and does not include figures, tables, or other visual elements. Incorporating these additional components could provide a more comprehensive assessment of LLMs' capabilities in the peer review process. Additionally, the dataset is currently limited to the NLP domain. It would be interesting to explore the performance of LLMs in other research areas. Expanding the dataset to include papers from various domains could help assess the generalizability of our findings and identify potential domain-specific challenges. Furthermore, our work focuses on the pre-rebuttal phase of the peer review process, assessing reviews based on the original submission. Incorporating the multi-turn aspect of peer review, including author rebuttals and post-rebuttal reviews, could offer a more comprehensive understanding of LLMs' capabilities in the entire review process.

References

AI@Meta. 2024. Llama 3 model card.

Signe Altmäe, Alberto Sola-Leyva, and Andres Salumets. 2023. Artificial intelligence in scientific writing: a friend or a foe? *Reproductive BioMedicine Online*, 47(1):3–9.

Anthropic. 2024. Introducing the next generation of claude.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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*, pages 1–23.

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.

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

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

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.

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*, pages 1–5.

Gemini Team Google. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

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

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Neha 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, Seattle, United States. 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.

Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. Multi-task peer-review score prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 121–126, Online. Association for Computational Linguistics.

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

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tribikram Pradhan, Chaitanya Bhatia, Prashant Kumar, and Sukomal Pal. 2021. A deep neural architecture based meta-review generation and final decision prediction of a scholarly article. *Neurocomputing*, 428:218–238.

Zafaryab Rasool, Stefanus Kurniawan, Sherwin Balugo, Scott Barnett, Rajesh Vasa, Courtney Chesser, Benjamin M Hampstead, Sylvie Belleville, Kon Mouzakis, and Alex Bahar-Fuchs. 2024. Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset. *Natural Language Processing Journal*, page 100083.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*.

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.

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, Dublin, Ireland. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, Portland, Oregon, USA. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

A Experiment Details

A.1 BERTScore

During the evaluation of LLMs' explanations for `Deficient` segments (Table 5, 10, and 11), we use `microsoft/deberta-xlarge-mnli` as the base model for computing BERTScore (Zhang et al., 2020), as officially suggested[1].

In the experiment of computing inter-LLM review complementarity (Section 4.1), we use `facebook/bart-large-mnli` as the base model for BERTScore. This is because `microsoft/deberta-xlarge-mnli` only supports input sequences up to 512 tokens, while some full reviews exceed this limit. In contrast, `facebook/bart-large-mnli` has a context size of 1024 tokens, making it suitable for processing longer reviews.

A.2 Similarity Score in `ITF-IDF`

We use SentenceBERT (Reimers and Gurevych, 2019) to calculate the similarity in `ITF-IDF`. We adopt the `all-MiniLM-L6-v2` pretrained model because it is fast and still offers good quality [2]. In practice, the similarity in our `ITF-IDF` can be computed using any sentence similarity model.

A.3 LLM Inference Details

Closed-source LLMs. We experiment with the following models and their corresponding API endpoints: GPT-4 (`gpt-4-turbo`), Gemini 1.5 (`gemini-1.5-flash-latest`), and Claude 3 (`claude-3-opus-20240229`).

Open-source LLMs. We experiment with the following models: Llama3-8B (`Meta-Llama-3-8B-Instruct`), Llama3-70B (`Meta-Llama-3-70B-Instruct`), and Qwen2-72B (`Qwen/Qwen2-7B-Instruct`).

For GPT-4, Claude 3, Gemini 1.5, and Qwen2-72B, we input the full prompt as shown in Table 12 and 13, which contains the complete instruction, paper title, full paper body text, and review text.

However, for Llama3-8B and Llama3-70B, the maximum supported context length is limited to 8k tokens[3]. To accommodate this constraint, we truncate the full paper body text while keeping the other components of the prompt intact. This is because the other components, such as the instruction, paper title, and review text, are crucial for the evaluation and cannot be truncated.

[1] https://github.com/Tiiiger/bert_score

[2] https://sbert.net/docs/sentence_transformer/pretrained_models.html

B Influence of Different Thresholds in `ITF-IDF`

Table 6 shows the impact of varying the similarity threshold t on the `ITF-IDF` scores on full reviews. The performance rank remains the same across different t values.

Model	$t = 0$	$t = 0.5$	$t = 0.7$	$t = 0.99$
GPT4	0.77	3.37	6.46	8.42
Claude3	0.87	4.09	7.61	9.32
Gemini	0.81	3.82	6.67	8.57
Human	1.22	6.04	8.45	9.50

Table 6: `ITF-IDF` under different t values. The rank remains the same across different t values.

C Error Types Detected by LLMs

Table 7 provide a statistics of the error types that LLMs successfully identify in the human-written reviews. We report the number and percentage of segments detected by each LLM for each error type.

D `Deficient` Segment Error Types

Table 9 present a comprehensive list of the error types used to categorize the `Deficient` segments in the reviews. Each error type is accompanied by an explanation defined by our annotation team. We also report the percentage of each error type for both human-written and LLM-generated reviews in Table 8.

E Explanation Score Across Different Prompts

This section compares the performance of LLMs in generating explanations for the correctly identified `Deficient` segments across different prompting strategies. We report the ROUGE and BERTScore values for each LLM and prompt combination in Table 10 and 11.

[3] https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Error Type	ReviewCritique	GPT4 / Claude Opus / Gemini 1.5	
		Identified	Percentage
Out-of-scope	31	14 / 21 / 25	45.2 / 67.7 / 80.6%
Inaccurate Summary	41	7 / 2 / 4	17.1 / 4.9 / 9.8%
Neglect	140	75 / 100 / 122	53.6 / 71.4 / 87.1%
Inexpert Statement	130	68 / 80 / 100	52.3 / 61.5 / 76.9%
Misunderstanding	163	77 / 111 / 120	47.2 / 68.1 / 73.6%
Vague Critique	66	39 / 52 / 57	59.1 / 78.8 / 86.4%
Misinterpret Novelty	27	19 / 23 / 22	70.4 / 85.2 / 81.5%
Misplaced Attributes	7	4 / 3 / 5	57.1 / 42.9 / 71.4%
Writing	20	2 / 2 / 4	10.0 / 10.0 / 20.0%
Superficial Review	19	2 / 2 / 3	10.5 / 10.5 / 15.8%
Invalid Criticism	20	11 / 12 / 16	55.0 / 60.0 / 80.0%
Invalid Reference	3	2 / 1 / 2	66.7 / 33.3 / 66.7%
Subjective	8	5 / 7 / 6	62.5 / 87.5 / 75.0%
Missing Reference	9	6 / 7 / 7	66.7 / 77.8 / 77.8%
Experiment	13	2 / 3 / 3	15.4 / 23.1 / 23.1%
Contradiction	5	1 / 1 / 1	20.0 / 20.0 / 20.0%
Summary Too Short	2	1 / 0 / 1	50.0 / 0.0 / 50.0%
Typo	2	1 / 2 / 2	50.0 / 100.0 / 100.0%
Concurrent Work	1	1 / 1 / 1	100.0 / 100.0 / 100.0%
Unstated Statement	2	0 / 0 / 0	0.0 / 0.0 / 0.0%
Copy-pasted Summary	2	0 / 0 / 1	0.0 / 0.0 / 50.0%
Misunderstanding Submission Rule	2	1 / 2 / 2	50.0 / 100.0 / 100.0%

Table 7: Comparison of GPT-4, Claude, and Gemini in identifying `Deficient` segments. Red-colored types have a significantly lower percentage compared to the average recall of LLMs.

F Prompt Templates

We provide the detailed prompt templates used for the experiments throughout the paper. This includes prompts for generating LLM reviews (Table 14) and identifying `Deficient` segments (Table 12 and 13).

Error Types	Human Review			LLM Review		
	All	**Acc.**	**Rej.**	**All**	**Acc.**	**Rej.**
Out-of-scope	4.35%	5.05%	3.79%	31.67%	34.91%	28.70%
Inaccurate Summary	5.75%	8.52%	3.54%	1.36%	0.94%	1.74%
Neglect	19.64%	24.29%	15.91%	4.52%	5.66%	3.48%
Inexpert Statement	18.23%	16.72%	19.44%	5.88%	4.72%	6.96%
Misunderstanding	22.86%	17.35%	27.27%	10.41%	9.43%	11.30%
Vague Critique	9.26%	5.99%	11.87%	7.69%	7.55%	7.83%
Misinterpret Novelty	3.79%	6.94%	1.26%	1.36%	1.89%	0.87%
Misplaced attributes	0.98%	0.95%	1.01%	-	-	-
Writing	2.81%	2.52%	3.03%	4.07%	2.83%	5.22%
Superficial Review	2.66%	3.15%	2.27%	9.95%	11.32%	8.70%
Invalid Criticism	2.81%	2.84%	2.78%	-	-	-
Invalid Reference	0.42%	0.32%	0.51%	-	-	-
Subjective	1.12%	1.89%	0.51%	-	-	-
Missing Reference	1.26%	0.63%	1.77%	-	-	-
Experiment	1.82%	1.89%	1.77%	1.36%	0.94%	1.74%
Contradiction	0.70%	-	1.26%	9.05%	8.49%	9.57%
Summary Too Short	0.28%	-	0.51%	-	-	-
Typo	0.28%	-	0.51%	-	-	-
Concurrent work	0.14%	-	0.25%	-	-	-
Unstated statement	0.28%	0.63%	-	7.69%	6.60%	8.70%
Copy-pasted Summary	0.28%	-	0.51%	-	-	-
Misunderstanding Submission Rule	0.28%	0.32%	0.25%	-	-	-
Duplication	-	-	-	4.98%	4.72%	5.22%

Table 8: Percentage fo error types in Human-written and LLM-generated reivews amaong all `Deficient` segments.

Error Type	Explanation
Misunderstanding	The reviewer misinterprets claims or ideas presented in the paper, leading to inaccurate or irrelevant comments.
Neglect	The reviewer overlooks important details explicitly stated in the paper, resulting in unwarranted questions or critiques.
Vague Critique	The review lacks specificity, claiming missing components without clearly identifying what is missing.
Inaccurate Summary	The summary in the review misrepresents the main content or contributions of the paper.
Out-of-scope	The reviewer suggests additional methods, experiments, or analyses that are beyond the intended scope of the paper.
Misunderstanding of the Submission Rule	The reviewer believes the submission format violates conference rules, but this is not actually the case.
Subjective	The review makes assertions about the paper's clarity or quality without providing sufficient justification or evidence.
Invalid Criticism	The reviewer's criticism is considered invalid, especially when suggesting impractical experiments or trivializing results.
Misinterpret Novelty	The reviewer questions the novelty of the work without substantiating their claims with relevant references
Superficial Review	The reviewer appears to have only skimmed the paper, providing generic or unsupported comments about the presence or absence of weaknesses.
Writing	Discrepancies arise when the reviewer praises the writing, while our annotator suggests it needs more clarity or explicitness.
Inexpert Statement	The reviewer exhibits a lack of domain knowledge, leading to unnecessary or irrelevant concerns.
Missing Reference	The reviewer proposes alternative frameworks or methods without providing justification or citing relevant references
Experiment	Conflicting opinions about the design of experiments; the reviewer praises them while our annotator suggests adding more baselines or tests.
Misplaced attributes	Strengths are incorrectly listed as weaknesses or vice versa.
Invalid Reference	The reviewer cites non-peer-reviewed sources or blogs, which is not appropriate for academic validation.
Unstated statement	Statements made in the review are not supported by content in the paper.
Summary Too Short	The provided summary is excessively brief, offering little to no insight into the actual content of the paper.
Contradiction	The reviewer contradicts themselves within the review, such as criticizing the paper's experiments while later stating that the experiments are comprehensive.
Typo	The review contains typographical errors that may affect clarity or understanding.
Copy-pasted Summary	The summary is directly copied from the submission.
Concurrent work	The reviewer requests comparisons with work conducted concurrently, which may not have been considered by the authors.
Duplication	The review segment is a repetition or duplication of a previous segment within the same review.

Table 9: Error types in paper reviews.

Model	ROUGE-1 / 2 / L / BERTScore	
	`Labeling-All`	`Select-Deficient`
GPT-4	16.12 / 2.05 / 13.58 / 56.87	17.13 / 2.71 / 14.64 / 55.63
Claude Opus	18.54 / 3.03 / 16.03 / 58.44	20.18 / 3.69 / 17.52 / 57.28
Gemini 1.5	19.40 / 2.99 / 17.14 / 58.10	18.47 / 2.98 / 16.38 / 56.46
Llama3-8B	15.97 / 1.74 / 14.14 / 56.23	16.49 / 2.22 / 13.65 / 55.23
Llama3-70B	15.03 / 2.25 / 13.04 / 58.19	15.94 / 1.95 / 13.78 / 57.09
Qwen2-72B	14.49 / 2.27 / 12.86 / 56.66	17.07 / 3.00 / 14.69 / 56.88

Table 10: Evaluation of LLMs' explanations for correctly identified `Deficient` segments with `Labeling-All` and `Select-Deficient` prompt methods.

Model	ROUGE-1 / 2 / L / BERTScore	
	Both "No"	Either "No"
GPT-4	16.79 / 2.46 / 14.16 / 56.21	16.61 / 2.36 / 14.09 / 56.25
Claude Opus	19.82 / 3.63 / 17.23 / 58.00	19.24 / 3.31 / 16.66 / 57.95
Gemini 1.5	19.25 / 3.08 / 17.12 / 57.42	18.88 / 2.99 / 16.72 / 57.17
Llama3-8B	16.94 / 2.22 / 14.49 / 56.07	16.17 / 1.91 / 13.92 / 55.86
Llama3-70B	15.72 / 2.02 / 13.63 / 57.64	15.44 / 2.12 / 13.38 / 57.71
Qwen2-72B	15.51 / 2.51 / 13.64 / 56.34	15.72 / 2.58 / 13.74 / 56.74

Table 11: Evaluation of LLMs' explanations for correctly identified `Deficient` segments with ensembling two prompts' results. The final scores are calculated by averaging the scores of each explanation generated by the prompts.

Assume you are a meta-reviewer of a natural language processing conference.

Given a paper submission and its corresponding review, your job is to assess the deficiency of each review segment.

The review is segmented, and each segment has an index at the start. You need to assess if each segment of the review is "deficient" or not

The criteria for “Deficient” are:

1. Sentences that contain factual errors or misinterpretations of the submission.

2. Sentences lacking constructive feedback.

3. Sentences that express overly subjective, emotional, or offensive judgments, such as “*I don’t like this work because it is written like by a middle school student*.”

4. Sentences that describe the downsides of the submission without supporting evidence, for example, “*This work misses some related work*.”

Your answer should be indexed according to the indices of the segments. For each segment, if it is "reliable," you can simply output "Yes." If it is `Deficient`, you should output "No," followed by the reason why it is `Deficient`.

In your assessment, consider not only the content of each segment but also the overall context of the review and the paper submission.

Here is the submission title:

{paper_title}

Here is the body text of the submission:

{body_text}

Here is the segmented review:

{review_text}

Here is the author rebuttals:

{author_rebuttals_text}

Your answer should only contain the segment index, your assessment "Yes" or "No," and the explanation if your assessment is "No." Here is an example format:

[index]. [Yes or No][Your explanation if your answer is No]

Output your answer below:

Table 12: `Labeling-All` prompt template.

Assume you are a meta-reviewer of a natural language processing conference. Given a paper submission and its corresponding review, your job is to assess the deficiency of each review segment.

The review is segmented, and each segment has an index at the start. You need to assess if each segment of the review is "deficient" or not

The criteria for “Deficient” are:

1. Sentences that contain factual errors or misinterpretations of the submission.

2. Sentences lacking constructive feedback.

3. Sentences that express overly subjective, emotional, or offensive judgments, such as “*I don’t like this work because it is written like by a middle school student*.”

4. Sentences that describe the downsides of the submission without supporting evidence, for example, “*This work misses some related work*.”

Your answer should include the indices of all `Deficient`segments, each followed by the reason why the segment is `Deficient`.

In your assessment, consider not only the content of each segment but also the overall context of the review and the paper submission.

Here is the submission title:

{paper_title}

Here is the body text of the submission:

{body_text}

Here is the segmented review:

{review_text}

Here is the author rebuttals:

{author_rebuttals_text}

Your answer should contain only the indices of all `Deficient`segments, followed by the reason why each segment is `Deficient`.

[index]. [Your explanation]

Output your answer below:

Table 13: `Select-Deficient` prompt template.

As an esteemed reviewer with expertise in the field of Natural Language Processing (NLP), you are asked to write a review for a scientific paper submitted for publication. Please follow the reviewer guidelines provided below to ensure a comprehensive and fair assessment:

Reviewer Guidelines: {review_guidelines}

In your review, you must cover the following aspects, adhering to the outlined guidelines:

Summary of the Paper: [Provide a concise summary of the paper, highlighting its main objectives, methodology, results, and conclusions.]

Strengths and Weaknesses: [Critically analyze the strengths and weaknesses of the paper. Consider the significance of the research question, the robustness of the methodology, and the relevance of the findings.]

Clarity, Quality, Novelty, and Reproducibility: [Evaluate the paper on its clarity of expression, overall quality of research, novelty of the contributions, and the potential for reproducibility by other researchers.]

Summary of the Review: [Offer a brief summary of your evaluation, encapsulating your overall impression of the paper.]

Correctness: [Assess the correctness of the paper's claims, you are only allowed to choose from the following options:

{Explanation on different correctness scores}

Technical Novelty and Significance: [Rate the technical novelty and significance of the paper's contributions, you are only allowed to choose from the following options:

{Explanation on different Technical Novelty and Significance scores}

Empirical Novelty and Significance: [Evaluate the empirical contributions, you are only allowed to choose from the following options:

{Explanation on different Empirical Novelty and Significance scores}

Flag for Ethics Review: Indicate whether the paper should undergo an ethics review [YES or NO].

Recommendation: [Provide your recommendation for the paper, you are only allowed to choose from the following options:

{Explanation on different recommendation scores}

Confidence: [Rate your confidence level in your assessment, you are only allowed to choose from the following options:

{Explanation on different confidence scores}

To assist in crafting your review, here are two examples from reviews of different papers:

Review Example 1:

{review_example_1}

Review Example 2:

{review_example_2}

Follow the instruction above, write a review for the paper below:

Table 14: Prompt template for generating reviews with LLMs