Automatically Evaluating the Paper Reviewing Capability of Large Language Models

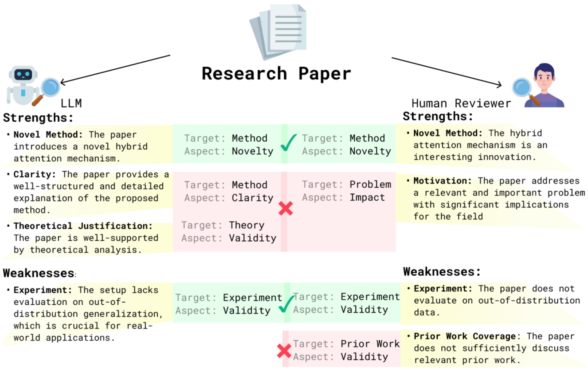
Anonymous ACL submission

Abstract

Peer review is essential for scientific progress, but it faces challenges such as reviewer shortages and growing workloads. Although Large Language Models (LLMs) show potential for providing assistance, research has reported significant limitations in the reviews they generate. While the insights are valuable, conducting the analysis is challenging due to the considerable time and effort required, especially given the rapid pace of LLM developments. To address the challenge, we developed an automatic evaluation pipeline to assess the LLMs' paper review capability by comparing them with expert-generated reviews. By constructing a dataset[1] consisting of 676 OpenReview papers, we examined the agreement between LLMs and experts in their strength and weakness identifications. The results showed that LLMs lack balanced perspectives, significantly overlook novelty assessment when criticizing, and produce poor acceptance decisions. Our automated pipeline enables a scalable evaluation of LLMs' paper review capability over time.

1 Introduction

Reviewing academic papers lies at the heart of scientific advancement, but it demands substantial expertise, time, and effort. The peer review system faces several challenges, including a growing number of submissions that outpace the reviewer availability, lack of incentives, and reviewer fatigue (Tropini et al., 2023; Horta and Jung, 2024; Hossain et al., 2025). While Large Language Models (LLMs) hold the potential to assist reviewers by reviewing papers automatically (Hosseini and Horbach, 2023; Robertson, 2023), prior research has reported significant limitations in their performance (Du et al., 2024; Liang et al., 2024; Zhou et al., 2024). For example, studies have highlighted that LLM-generated reviews often lack accuracy, detail, and specificity when compared to reviews written by human experts (Zhou et al., 2024; Mostafapour et al., 2024).



- **Novel Method:** The paper introduces a novel hybrid attention mechanism.
- **Clarity:** The paper provides a well-structured and detailed explanation of the proposed method.
- **Theoretical Justification:** The paper is well-supported by theoretical analysis.
- **Experiment:** The setup lacks evaluation on out-of-distribution generalization, which is crucial for real-world applications.

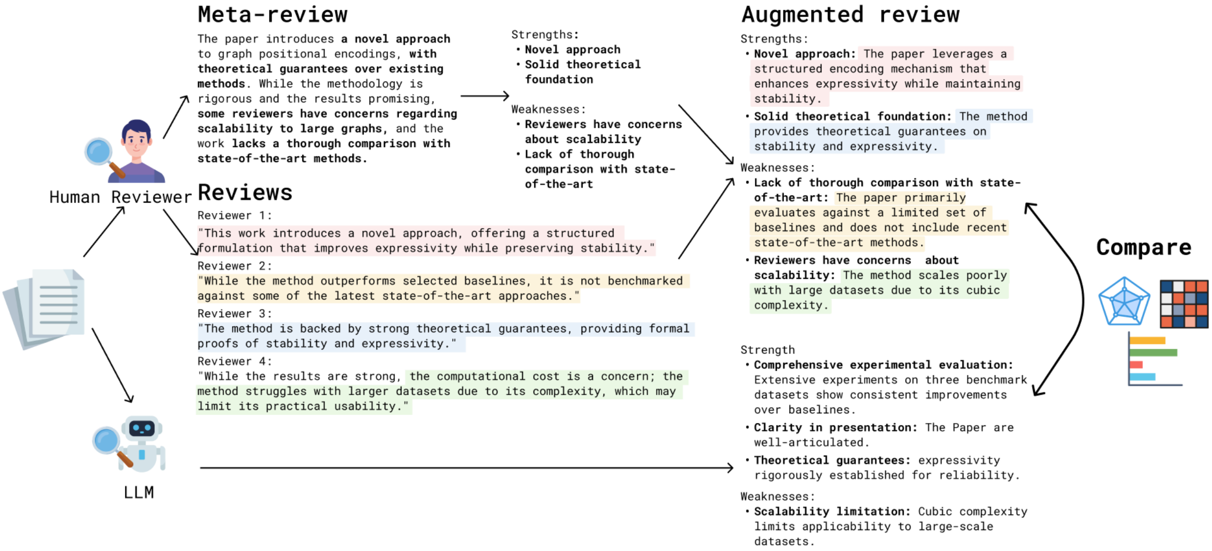
- **Novel Method:** The hybrid attention mechanism is an interesting innovation.
- **Motivation:** The paper addresses a relevant and important problem with significant implications for the field
- **Experiment:** The paper does not evaluate on out-of-distribution data.
- **Prior Work Coverage:** The paper does not sufficiently discuss relevant prior work.

Figure 1: We consider paper review task that generates a summary of paper, strengths, weaknesses, and the final judgement. Our goal is to examine the level of agreement between LLMs and human experts in reviewing papers, based on their feedback targets and aspects.

While these findings are informative, they do not sufficiently clarify the precise differences between expert-generated reviews and LLM-generated reviews. Specifically, it is still unclear to what extent LLMs excel or fall short in different aspects of reviewing compared to human experts. Addressing the question requires systematic quantitative analysis of review data, but such analysis is challenging to scale due to the significant time and effort required from researchers.

To address this gap, we introduce an automated pipeline designed to systematically analyze these differences. Our approach is to automatically annotate the strengths and weaknesses identified in reviews based on targets (e.g., problem, methodology, and experiment) and their associated aspects (e.g., validity, clarity, and novelty) and examine the agreement between LLMs and human experts (Figure 1). By introducing a systematic framework for examining the strengths and limitations of LLMs in academic review, this work offers valuable insights

[1] https://figshare.com/s/d5adf26c802527dd0f62



Meta-review

The paper introduces **a novel approach** to graph positional encodings, **with theoretical guarantees over existing methods**. While the methodology is rigorous and the results promising, **some reviewers have concerns regarding scalability to large graphs,** and the work **lacks a thorough comparison with state-of-the-art methods.**

Strengths:
- **Novel approach**
- **Solid theoretical foundation**

Weaknesses:
- **Reviewers have concerns about scalability**
- **Lack of thorough comparison with state-of-the-art**

Reviews

Reviewer 1:
"This work introduces a novel approach, offering a structured formulation that improves expressivity while preserving stability."

Reviewer 2:
"While the method outperforms selected baselines, it is not benchmarked against some of the latest state-of-the-art approaches."

Reviewer 3:
"The method is backed by strong theoretical guarantees, providing formal proofs of stability and expressivity."

Reviewer 4:
"While the results are strong, the computational cost is a concern; the method struggles with larger datasets due to its complexity, which may limit its practical usability."

Augmented review

Strengths:
- **Novel approach:** The paper leverages a structured encoding mechanism that enhances expressivity while maintaining stability.
- **Solid theoretical foundation:** The method provides theoretical guarantees on stability and expressivity.

Weaknesses:
- **Lack of thorough comparison with state-of-the-art:** The paper primarily evaluates against a limited set of baselines and does not include recent state-of-the-art methods.
- **Reviewers have concerns about scalability:** The method scales poorly with large datasets due to its cubic complexity.

Strength
- **Comprehensive experimental evaluation:** Extensive experiments on three benchmark datasets show consistent improvements over baselines.
- **Clarity in presentation:** The Paper are well-articulated.
- **Theoretical guarantees:** expressivity rigorously established for reliability.

Weaknesses:
- **Scalability limitation:** Cubic complexity limits applicability to large-scale datasets.

Figure 2: The overall evaluation process. Given a paper, we extract strengths and weaknesses from review data on the OpenReview platform. To identify key strengths and weaknesses that influence the final acceptance, we extracted them from the meta-review and augmented details from reviewer comments to make them self-contained. We then compared these with LLM-identified strengths and weaknesses, based on their feedback target and aspect. The evaluation is conducted automatically, enabling a scalable and longitudinal evaluation over time.

into improving their performance and enhancing their potential role in assisting the review process.

Our study leverages a dataset of 676 papers and their review data that has been collected from OpenReview[2] for ICLR conferences spanning 2021 to 2024. We extracted the strengths and weaknesses highlighted in the meta-reviews using gpt-4o[3], which will be compared with LLM-identified strengths and weaknesses. Then, we developed a coding schema (Table 1) for annotating the targets and aspects of these strengths and weaknesses, by surveying 9 AI paper submission guidelines and prior research on review analysis (Chakraborty et al., 2020; Ghosal et al., 2022; Yuan et al., 2022). Based on the schema, we manually annotated 327 strengths and weaknesses of 68 randomly sampled papers, providing a basis for building an automatic annotation tool. Our LLM-powered automatic annotation tool achieved 0.85 (target) and 0.86 (aspect) inter-rater reliability (IRR) with human-annotated results, showing high level of consistency and accuracy. Overall, we identified 3,657 review items (1,231 strengths and 2,416 weaknesses) from the review dataset.

We evaluated 8 LLMs (4 GPT, 2 Llama, and 2 DeepSeek family) for their paper review capability. After generating reviews for each of the 676 papers, we analyzed the agreement between LLMs' and experts' reviews based on their target and aspect assigned to strengths and weaknesses. The results showed that: 1) LLMs lack balanced perspective compard to human experts, 2) LLMs significantly neglect novelty assessment for evaluating papers' weaknesses, and 3) the paper acceptance decisions are not accurate. The findings are consistent for all the LLMs, highlighting clear opportunities for improving their reasoning capability.

We release a dataset comprising 68 papers, experts reviews, 3,657 strengths and weaknesses identified from the reviews with automatically annotated targets and aspects, LLM-generated reviews from 8 LLMs, and a total of 43,042 strengths and weaknesses identified from the LLMs with their annotated targets and aspects.

2 Task

We define the paper review generation task as follows: given a research paper, 1) summarize main points, 2) identify a list of strengths and weaknesses, and 3) predict the final acceptance of the paper. This task offers direct value to various user groups (e.g., authors who want to get initial feedback on their draft, or reviewers who want to examine diverse view points) by providing actionable feedback for improving their papers. While valuable, evaluating papers based on research standards (e.g., novelty, rigor, and clarity) is difficult for LLMs as it requires significant expertise.

[2] https://openreview.net/
[3] gpt-4o-2024-08-06

3 Constructing Expert Review Dataset

3.1 Collecting Review Data

We used real-world review data covering ICLR 2021-2024 from the OpenReview platform[4], where human experts evaluated submissions for a top-tier AI conference. Using the OpenReview API[5] and the list of submissions from public GitHub repositories[6], we initially collected 18,407 submissions with their review data.

3.2 Identifying Strengths and Weaknesses

One of the challenges in identifying the strengths and weaknesses of these papers is that each review consists of multiple blocks, including a meta-review and individual review texts from several reviewers. To address the challenge, our approach is to use meta-review, a final review from a qualified expert that summarizes reviews and highlights important strengths and weaknesses for supporting the final decision. As the meta-review does not capture all the details, we created self-contained strengths and weaknesses by 1) extracting them from the meta-review and 2) augmenting these extracted elements with detailed comments from individual reviews (non-meta). We designed a prompting chain that consists of three prompts (Appendix A.1.1). After excluding withdrawn submissions that lack meta-reviews, 14,922 submissions remained.

4 Building an Automatic Annotator Based on Target and Aspect

The central goal of this paper is to analyze where LLMs excel and fall short in reviewing papers, compared to human experts. To achieve the goal, we 1) annotate each of the strengths and weaknesses identified by LLMs and experts and 2) examine the agreement between them based on the annotation results. The analysis offers insights into the distinct contributions and limitations of LLMs in reviewing papers, informing strategies to foster more effective human-LLM collaboration in reviewing papers.

4.1 Developing a Coding Scheme

Our focus in classifying strengths and weaknesses lies in two key dimensions: targets (i.e., what the review praises or critiques) and aspects (i.e., the specific elements of the target being evaluated). To build an initial codebook, we surveyed 9 AI paper submission guidelines (Appendix A.2.1) and extracted target-aspect pairs from each statement in the guidelines (e.g., *"The paper should state the full set of assumptions of all theoretical results if the paper includes theoretical results."* yields the target *Theory* and aspect *Completeness*). We also reviewed related work on the analysis of paper review data (Chakraborty et al., 2020; Ghosal et al., 2022; Yuan et al., 2022). After identifying 33 targets and 13 aspects, we merged similar items to create simple and distinct categories, resulting in 7 targets and 4 aspects. Table 1 shows our final coding scheme.

4.2 Building an Automatic Classifier Based on Human Annotation

Based on the coding scheme, we annotated targets and aspects of strengths and weaknesses to produce ground truth for developing an automatic annotator. We randomly sampled 68 papers from our review dataset, yielding 327 instances of strengths and weaknesses. Two authors annotated each instance together, resolving any conflicts. Most conflicts arose when an instance illustrated multiple points. For example, an instance such as "***Technically sound with a strong foundation**: The paper's technical foundation is evident in its bi-level optimization framework, ... Technical novelty also arises from using supermartingale constraints on the barrier function ...*" could correspond to both *Validity* and *Novelty* aspect. Two authors finalized the annotation through discussions, focusing on the main point or root cause of the issue. In the example, we annotated *Validity*, as the strength mainly praises the technical soundness, as shown in the header.

We then designed prompts to automatically annotate the instances, assigning a target and aspect label to each. Specifically, we designed four prompts where each corresponds to one of the four combinations of target/aspect and strength/weakness A.2.2. Table 2 shows the inter-rater reliability (IRR) between author annotations and LLM annotations. Classification using `o3-mini` achieved the IRR scores of 0.85 for targets and 0.86 for aspects. Given the high IRR and its relatively low computational cost, we used `o3-mini` for the automatic annotation of both target and aspect in the main evaluation. Moreover, an examination of the con-

[4] The review data is publicly available and permits use of data for research.

[5] https://docs.openreview.net/getting-started/using-the-api

[6] https://github.com/{evanzd/ICLR2021-OpenReviewData, fedebotu/ICLR2022-OpenReviewData, fedebotu/ICLR2023-OpenReviewData, hughplay/ICLR2024-OpenReviewData}

Target	
Code	**Definition (The review addresses ...)**
Problem	Motivation, task definitions, and problem statements.
Prior Research	References and contextual positioning of the submission.
Method	Proposed approach, techniques, algorithms, or datasets.
Theory	Theoretical foundations, assumptions, proofs, or justifications.
Experiment	Experimental setup, results, and analysis.
Conclusion	Findings, implications, discussions, and takeaways.
Paper	General targets of the paper without specifying a particular target
Aspect	
Code	**Definition (The review addresses ...)**
Impact	Significance or practical influence of the work.
Novelty	Originality of the submission compared to prior research.
Clarity	Readability, ambiguity, or communication aspects.
Validity	Soundness, completeness, and rigor.
Not-specific	Multiple targets without emphasis on a particular aspect.

Table 1: The coding schema. To identify codes for targets (i.e., what the review praises or critiques) and aspects (i.e., the specific elements of the target being evaluated), we surveyed 9 AI paper submission guidelines (Appendix A.2.1) and prior research on review analysis (Chakraborty et al., 2020; Ghosal et al., 2022; Yuan et al., 2022).

fusion matrix (Appendix A.2.3) suggests that the errors tend to occur in semantically related categories, indicating that the misclassifications are not arbitrary but rather reflect subtle ambiguities inherent in the data.

Model	Target	Aspect
gpt-4o-mini	0.75	0.80
gpt-4o	0.87	0.83
o3-mini	0.85	0.86

Table 2: Inter-Rater Reliability between annotations of authors and LLMs for the target and aspect.

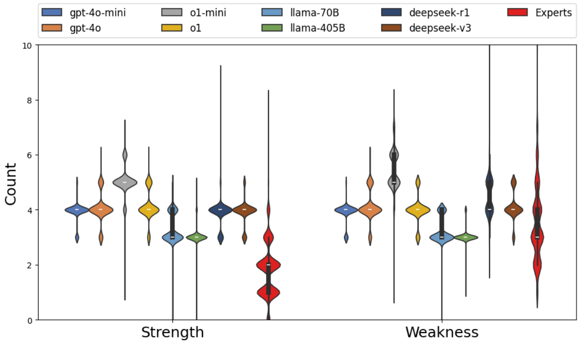


Figure 3: Distribution of strengths and weaknesses. Unlike human experts, LLMs reported a consistent count regardless of paper contents. o1-mini identified the most, while Llama models identified the fewest points.

5 Evaluation

The goal of our evaluation is to analyze the strengths and weaknesses of a given paper as identified by LLMs, comparing them with those identified by human experts. Note that our evaluation does not consider the correctness of the identified strengths and weaknesses because our focus is comparing perspectives in reviewing papers for both groups, not the content itself.

The evaluation is based on paper-review pairs. However, we excluded *accepted* submissions in the evaluation because OpenReview provides the camera-ready versions (post-review) rather than the submitted versions (pre-review), leading to a mismatch between the collected review and the camera-ready paper. Therefore, we only focused on *rejected* papers, where the meta-review corresponds to the latest version of the paper. Out of 9,139 rejected papers, we randomly sampled 7.5% of them (685 papers) for the evaluation. In total, we obtained 3,689 review items (1,241 strengths and 2,448 weaknesses), each automatically annotated with a target and aspect label.

5.1 Large Language Models

We consider eight off-the-shelf LLMs, differing in size and availability (open-source vs. proprietary): four GPT models (gpt-4o-mini, gpt-4o, o1-mini,

	Overall			Strength			Weakness		
Model	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
DeepSeek-R1	**0.373**	0.314	0.460	0.341	0.254	0.520	**0.400**	**0.379**	0.424
o1-mini	0.359	0.283	**0.491**	0.331	0.232	**0.578**	0.385	0.343	**0.439**
o1	0.355	0.300	0.436	0.318	0.234	0.495	0.388	0.377	0.400
DeepSeek-V3	0.351	0.300	0.421	0.330	0.246	0.501	0.368	0.362	0.374
Llama-405B	0.350	**0.323**	0.381	**0.349**	**0.279**	0.465	0.350	0.371	0.331
gpt-4o	0.349	0.287	0.442	0.342	0.252	0.533	0.354	0.325	0.388
gpt-4o-mini	0.344	0.289	0.427	0.335	0.246	0.522	0.353	0.337	0.369
Llama-70B	0.339	0.302	0.388	0.338	0.260	0.481	0.341	0.350	0.332

Table 3: Overall performance of alignments on strengths and weaknesses between experts-identified and LLM-identified reviews. The metrics were computed by comparing the (target, aspect) set between experts' and LLMs' review. DeepSeek-R1 achieved the best performance, o1-mini achieved superior recall, and Llama-405B achieved superior precision, compared to other models.

o3-mini, o1)[7], two Llama models (Llama 3.1-{70B, 405B}), and two DeepSeek models (DeepSeek-{V3, R1}). We used the default parameters of the models.

5.2 Procedure

For each of the 685 papers, we generated review data using the prompting pipeline (Section 3), extracted the strengths and weaknesses, and annotated the corresponding targets and aspects using the automatic annotator powered by o3-mini. Then for each LLM in Section 5.1, we generated a review for each paper (See Appendix A.1.2 for the prompt), extracted the strengths and weaknesses, and annotated the targets and aspects using the same automatic annotator. Then we compared the annotated targets and aspects between the experts' reviews and LLMs' reviews.

5.3 Result

While human experts raised various number of points, LLMs identified a relatively consistent number of points regardless of the paper's content. Moreover, LLMs identified a similar number of points between strengths and weaknesses, which was a different pattern from that of the human experts. Figure 3 shows the distribution of the number of strengths and weaknesses identified by human experts and LLMs. Overall, LLMs identified more points on average (7.88) than human experts (5.39). Among the LLMs, Llama models identified fewer (3.17 strengths and 3.15 weaknesses, on average) whereas o1-mini reported more strengths and weaknesses (5.03 and 5.47, respectively) than other models. By comparing target and aspect labels between human experts and LLMs, we report the following key findings.

Overall, LLMs do not effectively identify key targets and aspects when reviewing papers. Table 3 shows the overall performance of LLMs, which computes the agreement of (target, aspect) labels between human experts and LLMs. The best F1 score among the LLMs was 0.37, which indicates a low level of agreement with human experts in identifying strengths and weaknesses. Since we only considered whether the categories of review items match rather than their detailed content, the result implies that the actual content of strengths and weaknesses would be significantly different between human experts and LLMs. In general, LLMs showed higher recall than precision scores, mainly due to the nature of identifying a higher number of review points than human experts. Also, LLMs consistently achieved higher F1 scores for weaknesses than strengths. Among the LLMs, deepseek-r1 achieved the best overall performance, o1-mini achieved the best recall, and Llama-405B achieved the best precision.

While overall agreement is low, both groups primarily emphasized technical validity and novelty in the strengths, and focused on technical validity and clarity in the weaknesses. Figure 4 shows the normalized distribution of target and aspect labels for both experts and LLMs. For targets, both groups primarily focused on core technical elements—Method, Experiment, and Theory. However, strengths and weaknesses illustrated different patterns: both groups praised Method more

[7] gpt-4o-2024-08-06, gpt-4o-mini-2024-07-18, o1-mini-2024-09-12, o1-2024-12-17

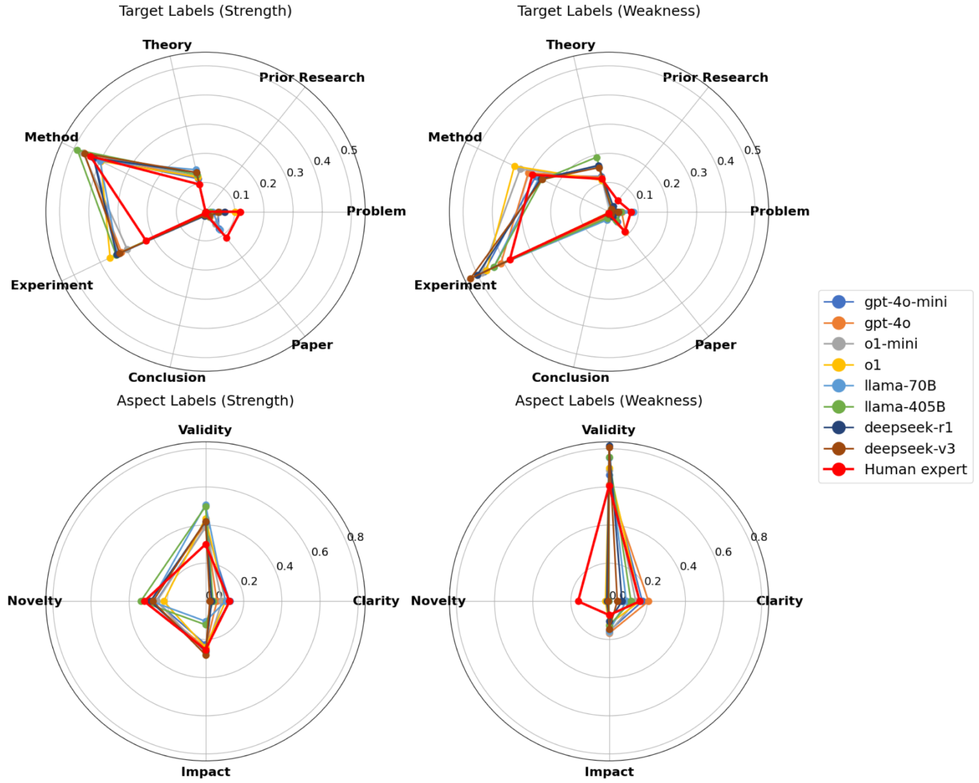


Figure 4: Normalized distributions by target/aspect and strength/weakness for LLMs and human experts (red line). Overall, both groups showed similar perspectives in reviewing papers, focusing on technical targets (i.e., Method, Experiment, and Theory) and validity. However, LLMs showed more biased perspectives that focus on the technical validity whereas human experts exhibited more balanced perspectives. However, all the LLMs lack consideration of Novelty for weaknesses compared to human experts, which is a significant limitation in reviewing papers.

than Experiment in the strengths, but criticized Experiment more than Method in the weaknesses. For aspects, both groups considered Validity as an important aspect, especially when evaluating weaknesses. Impact received more attention than Clarity in the strengths, whereas the opposite was observed in the weaknesses.

LLMs *consistently* exhibited a more biased perspective, while human experts maintained a more balanced perspective. Although both groups shared a core perspective, LLMs tend to focus on a few specific dimensions. For instance, LLMs focused primarily on Method and Experiment, while neglecting Prior Research (e.g., whether the paper adequately addresses prior work in positioning) and Problem (e.g., whether the task needs community attention), which human experts point out (Problem in the strengths and Prior Research in the weaknesses). For aspects, LLMs mostly focused on Validity in both strengths and weaknesses. In contrast, human experts considered the aspects more evenly among Validity, Novelty, and Clarity. Notably, LLMs exhibited a significant bias for Novelty. LLMs praised Novelty in the strengths, whereas they rarely criticized it in the weaknesses. This is a significant drawback, as a paper review requires a critical examination of novelty, by comparing them against existing work.

Due to their biased focus, the level of agreement between LLMs and human experts varied across different labels. Table 4 shows F1 scores for specific targets and aspects. For targets and aspects that LLMs focus more on — Method and Experiment targets and Validity aspect — LLMs had a much higher level of agreement with human experts compared to other targets and aspects. In the case of Experiment, the F1 score was consistently higher for weaknesses than strengths, suggesting that LLMs are more effective at identifying concerns in experiments (e.g., lack of baselines or scope of evaluation) than recognizing strong points of theories (e.g., experiments are rigorous and thorough). Similarly, for aspects other than Validity, agreement levels were notably lower. In particular, Novelty in the weaknesses, which LLMs largely overlooked, showed a significantly lower F1 score.

Table 4: F1 Score for target and aspects between DeepSeek-R1, o1-mini, and Llama-405B. Due to the biased perspective of LLMs, we observed a clear gap between what LLMs mostly focus on (e.g., Method and Experiment targets and Validity aspect) and overlook (e.g., Problem target and Novelty aspect). Full results (F1 score, precision, and recall across models and target/aspect labels) are available in Appendix A.2.4.

Target F1 score

Target	DeepSeek-R1	o1-mini	Llama-405B
Problem	0.30 0.40 / 0.20	0.28 0.35 / 0.20	0.16 0.16 / 0.16
Method	**0.73** 0.75 / 0.71	**0.76** 0.75 / 0.77	**0.69** 0.76 / 0.63
Theory	0.47 0.44 / 0.51	0.47 0.41 / 0.53	0.43 0.46 / 0.40
Experiment	**0.68** 0.51 / **0.85**	**0.68** 0.51 / **0.86**	**0.66** 0.52 / **0.81**

Aspect F1 score

Aspect	DeepSeek-R1	o1-mini	Llama-405B
Novelty	0.39 0.66 / **0.12**	0.39 0.66 / **0.12**	0.34 0.66 / **0.01**
Impact	0.41 0.54 / 0.29	0.43 0.56 / 0.30	0.32 0.35 / 0.29
Validity	**0.77** 0.60 / 0.95	**0.77** 0.60 / 0.95	**0.77** 0.60 / 0.95
Clarity	0.27 0.17 / 0.36	0.40 0.30 / 0.50	0.28 0.16 / 0.40

LLMs showed similar patterns in their emphasis in reviewing papers, regardless of their size and reasoning capability. All LLMs, including both proprietary and open source models, showed similar patterns that focused primarily on technical (Method, Experiment, and Theory) validity than on Novelty for the weaknesses. This consistency indicates that the observed biases could stem from the inherent design and training methods of LLMs, revealing potential room for improvement in the reasoning capability that requires leveraging external information (e.g., identifying comparable related work and analyzing novelty of submissions).

Final acceptance decisions are not accurate. Table 5 shows the rejection rate reported by each LLM. Overall, the best achieved rejection rate was 24.9%, which indicates that recall for rejected papers is poor. gpt-4o, Llama-405B, and DeepSeek-R1 performed significantly better than other models. gpt-4o and Llama-405B showed relatively high rejection rates while their agreement with human experts on strengths and weaknesses was low (0.348 and 0.349). DeepSeek-R1 demonstrated a moderate rejection rate among the models, with a relatively higher agreement score on strengths and weaknesses (0.373).

6 Discussion

In this paper, we found gaps in the way of reviewing papers between human experts and LLMs and reported several limitations of LLMs as an automated reviewer, using an automated pipeline. Based on the results, we discuss the following implications.

Table 5: Rejection percentage by model. 100% is the highest score, as we considered *rejected* papers. All the models were highly positive about the paper acceptance, although the papers were rejected.

Model	Rejection (%)
gpt-4o	27.92%
Llama-405B	24.30%
DeepSeek-r1	23.79%
gpt-4o-mini	9.50%
DeepSeek-v3	7.93%
o1-mini	5.45%
o1	3.36%
Llama-70B	0.74%

There exists significant room for improving alignments between human experts and LLMs in reviewing papers. Our results show that LLMs exhibit a more biased perspective, which mostly examines technical validity without contextual consideration, compared to human experts. To reduce the gap, fine-tuning models using our dataset could serve as a starting point. While our results revealed significant limitations of LLMs in reviewing papers, our focus was mostly on the target and aspect labels rather than comparing actual content. We suspect that a more significant gap lies in the actual content addressed in the review items, even if they share the same target and aspect labels. For instance, (Experiment, Validity) could point out either lack of necessary baselines or lack of ablation studies to justify authors' arguments. Content-level investigations may reveal more limitations of LLMs, ultimately contributing to improving the reasoning capability of LLMs.

Research should investigate the task of assessing the novelty of academic papers. Our finding illustrated that all LLMs in our analysis significantly overlooked the novelty aspect when evaluating weaknesses of papers. Previous studies have indicated that language models' ability to assess novelty is inferior to that of experts (Julian Just and Hutter, 2024; Lin et al., 2024), emphasizing the need to encourage LLMs to focus on novelty evaluation. Although novelty is one of the most important aspects in reviewing papers and efforts have been made to enhance LLMs' ability to assess novelty (Bougie and Watanabe, 2024; Lin et al., 2024), there exists no suitable benchmark for systematically measuring the novelty assessment capability of LLMs. We believe that creating the benchmark is a valuable contribution to the field, which allows LLMs to learn how to assess similarities between papers. Leveraging data in OpenReview could be an initial step as it contains experts' judgement on novelty of the paper for both positive and negative decisions.

It is important to further explore the alignment between points of strengths and weaknesses and the final decision based on them. We found that there exists a discrepancy between achieving a high level of agreement in strengths and weaknesses and correctly predicting the final decision. It implies that the way LLMs make the final decision based on the identified strengths and weaknesses could be different from the way of human experts. Consistent with our findings, previous research revealed that LLMs offer positive assessments (Latona et al., 2024a). As it is important to inform clear and convincing rationale behind the final decision, further investigation is needed to carefully evaluate whether the final decision based on the strengths and weaknesses is reasonable, for various stakeholders such as domain experts or novices. Generating reviews with clear alignments between the review points and final decision is a challenging task because often the relationship is not very clear and implicit (Zhou et al., 2024), relying on community norms and social factors. Learning the relationship from the review data could be useful for understanding the gap between human experts and LLMs in decision-making.

7 Related Work

With the powerful reasoning capability of LLMs, LLMs have the potential to assist in the task of reviewing papers (Latona et al., 2024b; D'Arcy et al., 2024). Research has explored the capability of LLMs in reviewing papers, identifying a set of limitations. While LLM-generated reviews can be helpful (Liang et al., 2024; Tyser et al., 2024; Lu et al., 2024), research has shown that LLMs-generated reviews lack diversity (Du et al., 2024; Liang et al., 2024) and technical details (Zhou et al., 2024), exhibit bias (Ye et al., 2024), tend to provide positive feedback (Zhou et al., 2024; Du et al., 2024), and may include irrelevant or even inaccurate comments (Mostafapour et al., 2024). Furthermore, research also has reported that LLM-generated reviews have a low level of agreement with experts-generated reviews (Saad et al., 2024).

To assess the quality of review, research has taken a quantitative approach by analyzing review text. For instance, research has evaluated the quality of review based on human preferences (Tyser et al., 2024), similarity to human-generated review (Zhou et al., 2024; Liang et al., 2024; Gao et al., 2024; Sun et al., 2024; Chamoun et al., 2024) and classification-based scores (Li et al., 2023). Another approach is to classify review data based on categories such as section (Ghosal et al., 2022), aspect (Yuan et al., 2022; Chamoun et al., 2024; Liang et al., 2024) and actionability (Choudhary et al., 2022). While quantitative approach provides concrete insights, it is typically conducted as a one-time evaluation, challenging to apply the consistent methodology over time.

8 Conclusion

We introduced an automatic evaluation pipeline to assess LLMs' capability in reviewing papers, by examining the agreement between LLM-generated and expert-generated reviews based on their target and aspect annotations for strengths and weaknesses. Our findings suggest that LLMs need to adopt a more balanced perspective, place greater emphasis on novelty assessment when critiquing papers, and better formulate their final judgement based on the identified strengths and weaknesses. We believe that our automated pipeline can contribute to ongoing evaluation of LLMs' paper review capabilities within the rapid pace of LLM developments, offering concrete insights for improving their reasoning capability.

Limitation

This paper has the following limitations. First, our dataset focuses soly on ICLR submissions and the coding schema is developed based on AI venues, which limit generalizability to other fields. Second, our analysis examines the target and aspect of the review items, but other important dimensions such as level of specificity and depth of justification remain unexplored. Third, while our automatic annotator achieved high IRR (0.85) with human annotations, some discrepancies still exist. Finally, we did not explore possible prompt engineering strategies that could mitigate the limitations of LLMs in paper review. Future work can investigate techniques to enhance the alignment between LLMs and human experts.

Ethical impact

This paper presents potential risks. First, while our vision is to build LLMs to effectively assist review process, our work could inadvertently encourage over-reliance on LLM-generated reviews among various user groups, including reviewers and novice researchers. Second, although our dataset could contribute to improving LLM performance of reviewing papers, it may introduce a certain bias due to the source of dataset; ICLR for papers and code based on AI research. Finally, we assess the quality of review based on alignment with expert reviews, but it could offer a potentially biased perspective, as our coding schema only considers two dimensions, which may undervalue the unique contributions of LLM-generated reviews.

References

Nicolas Bougie and Narimasa Watanabe. 2024. Generative adversarial reviews: When llms become the critic. *ArXiv*, abs/2412.10415.

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

Eric Chamoun, Michael Schlichktrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. *arXiv preprint arXiv:2405.20477*.

G. Choudhary, Natwar Modani, and Nitish Maurya. 2022. React: A review comment dataset for actionability (and more). *ArXiv*, abs/2210.00443.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. *arXiv preprint arXiv:2406.16253*.

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

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one*, 17(1):e0259238.

Hugo Horta and Jisun Jung. 2024. The crisis of peer review: Part of the evolution of science. *Higher Education Quarterly*, page e12511.

Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Md. Mahadi Hassan, Matthew Freestone, Matthew C. Williams Jr., Dongji Feng, and Santu Karmaker. 2025. Llms as meta-reviewers' assistants: A case study. *Preprint*, arXiv:2402.15589.

Mohammad Hosseini and Serge P.J.M. Horbach. 2023. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8.

Johann Füller Julian Just, Thomas Ströhle and Katja Hutter. 2024. Ai-based novelty detection in crowdsourced idea spaces. *Innovation*, 26(3):359–386.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. 2024a. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Preprint*, arXiv:2405.02150.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. 2024b. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*.

Miao Li, Eduard H. Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Conference on Empirical Methods in Natural Language Processing*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Ethan Lin, Zhiyuan Peng, and Yi Fang. 2024. Evaluating and enhancing large language models for novelty assessment in scholarly publications. *ArXiv*, abs/2409.16605.

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

Mehrnaz Mostafapour, Jacqueline H Fortier, Karen Pacheco, Heather Murray, and Gary Garber. 2024. Evaluating literature reviews conducted by humans versus chatgpt: Comparative study. *Jmir ai*, 3:e56537.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *ArXiv*, abs/2307.05492.

Ahmed Saad, Nathan Jenko, Sisith Ariyaratne, Nick Birch, Karthikeyan P Iyengar, Arthur Mark Davies, Raju Vaishya, and Rajesh Botchu. 2024. Exploring the potential of chatgpt in the peer review process: an observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 18(2):102946.

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

Carolina Tropini, B Brett Finlay, Mark Nichter, Melissa K Melby, Jessica L Metcalf, Maria Gloria Dominguez-Bello, Liping Zhao, Margaret J McFall-Ngai, Naama Geva-Zatorsky, Katherine R Amato, et al. 2023. Time to rethink academic publishing: the peer reviewer crisis.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. 2024. Ai-driven review systems: evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhen fei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *ArXiv*, abs/2412.01708.

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

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 Appendix

A.1 Review Generation

A.1.1 Prompts for Expert Review Generation

In this section, we provide prompts for identifying key strength and weakness from review data. Figure 5 shows the prompt for extracting weakness and strength from meta-review. Figure 6 shows the prompt for using detailed comments from reviews to augment the extracted elements. Figure 7 shows the prompt for removing some extraneous reference. We used the three prompts in a prompt chain, sequentially running the prompts.

Prompt for Meta-Review Summarization

```
[[ Meta-review ]]
%s

[[ Instruction ]]
Restructure the meta-review by (1) summary of the paper, (2) strengths, (3) weaknesses, and (4) final judgement.
Strengths and weaknesses should be in bullet points. Make sure that you do not paraphrase the original text but write
them as is as much as possible.

First, describe what the meta-review describes for each of the four points.
Second, restructure the meta-review by the four points.

[[ Your Response ]]

# What meta-review describes for each of the four points

# Restructured meta-review, preserving the original text as much as possible

## Summary of the paper

## Strengths

## Weaknesses

## Final judgement
```

Figure 5: Prompt for Meta-Review Summarization

Prompt for Generating Augmented Review

```
%s

[[ Instruction ]]

Refering to the reviews, add details on each bullet point in the meta-review's strengths and weaknesses. Make sure that you include (1) headers for each bullet point and (2) sufficient details for each bullet point from the reviews so that the meta-review's strengths and weaknesses are complete and comprehensive.

First, for each bullet point in below reflection, explain which additional details have been discussed in the reviews. Do not revise the bullet point contents. Discuss the details for each of the reviews separately. Make sure that you include sufficient details mentioned in the reviews such as numbers and technical terms so that the details provide concrete strengths and weaknesses.
Second, you are a senior reviewer who needs to write complete, logical, and self-contained meta-review, based on your explanation. Make sure that your strengths and weaknesses bullet points should be exactly the same with your reflection. Also, make sure that your strength and weakness bullet points with headers, capturing the reviewer comments in a complete manner. You may want to have multiple sentences for each header to comprehensively capture the reviewer comments. Do not refer to "reviewers" because you are writing your review, but writing the review in a very specific and concrete manner, including important numbers and technical terms.

# Reflection of strengths and weaknesses in the restructured meta-review

%s

[[ Your Response ]]

# Additional details from the reviews for each bullet point in the reflection where headers remain unchanged

# Complete, logical, and self-contained meta-review where strengths and weaknesses bullet points are exactly the same with that of the reflection

## Summary of the paper

## Strengths

## Weaknesses

## Final judgement
```

Figure 6: Prompt for Generating Augmented Review

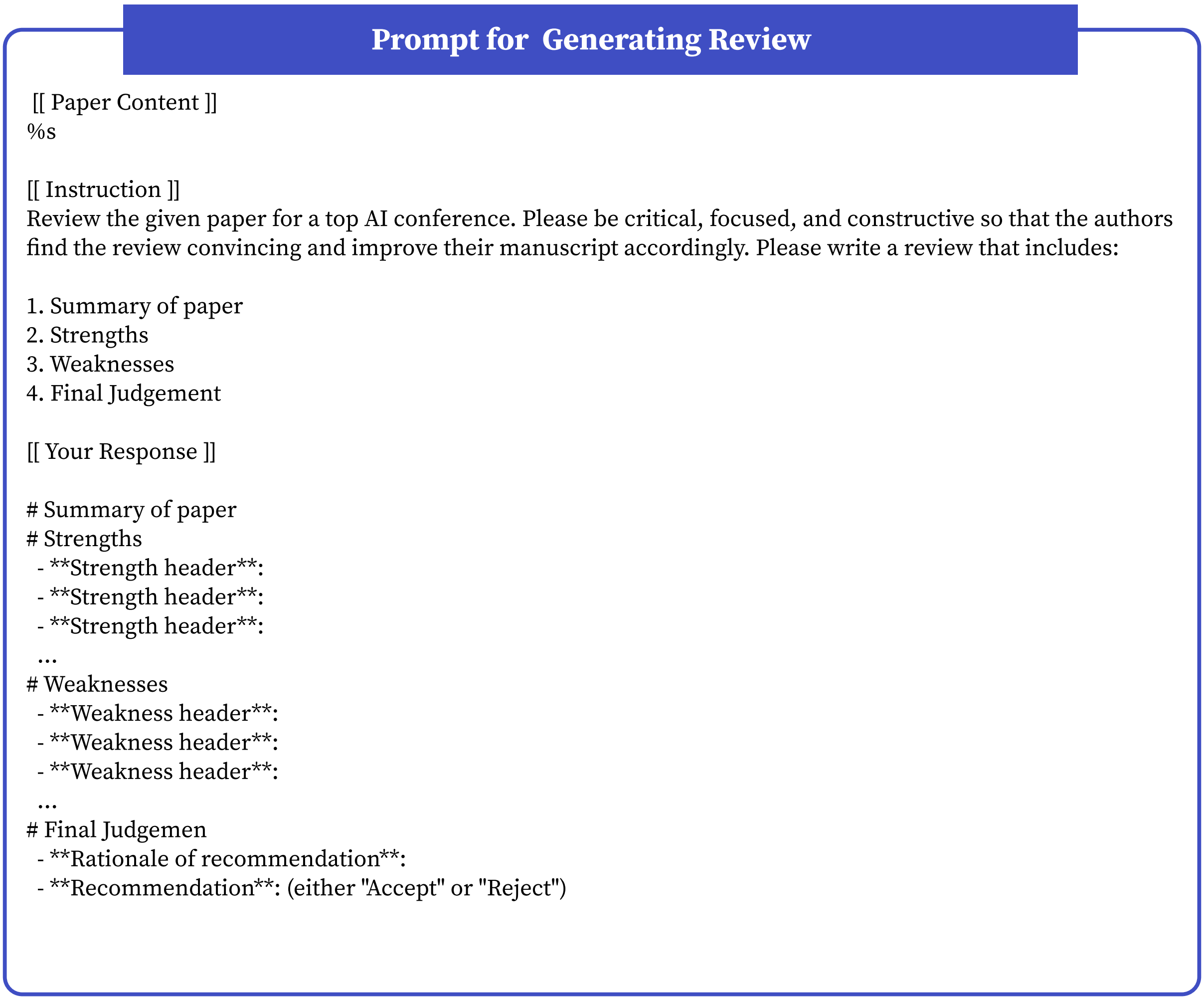
Prompt for Paraphrasing Augmented Review

```
[[ Review ]]

%s

[[ Instruction ]]

Given the "Review", paraphrase the **headers** of bullet points in the strengths and weaknesses so that the headers effectively summarizes the contents. Make sure that their body texts remain unchanged as much as possible, but paraphrase the body text minimally to remove any "reviewer" information such as reviewer's id or referencing reviewers as third person, just for that case. Also, make sure to attach "Summary of the paper" and "Final judgement" as exactly the same as in the "Review".

[[ Your Response ]]

## Summary of the paper

## Strengths

## Weaknesses

## Final judgement

# augment_review_template =
```

Figure 7: Prompt for Paraphrasing Augmented Review

A.1.2 Prompts for LLM Review Generation

Figure 8 shows the prompt for using LLM to generate reviews from paper.



Prompt for Generating Review

```
 [[ Paper Content ]]
%s

[[ Instruction ]]
Review the given paper for a top AI conference. Please be critical, focused, and constructive so that the authors
find the review convincing and improve their manuscript accordingly. Please write a review that includes:

1. Summary of paper
2. Strengths
3. Weaknesses
4. Final Judgement

[[ Your Response ]]

# Summary of paper
# Strengths
 - **Strength header**:
 - **Strength header**:
 - **Strength header**:
 ...
# Weaknesses
 - **Weakness header**:
 - **Weakness header**:
 - **Weakness header**:
 ...
# Final Judgemen
 - **Rationale of recommendation**:
 - **Recommendation**: (either "Accept" or "Reject")
```

Figure 8: Prompt for LLM Review Generation

A.2 Details of Building Automatic Annotator

A.2.1 AI paper writing guidelines

To ensure guidelines are comprehensive, we collected guidelines from 9 sources, comprising a total of 243 items, as shown in Table 6. An item refers to a specific requirement mentioned in the guidelines, which serves as a distinct criterion for reviewing or writing a paper.

Table 6: Guidelines and Item Count Summary

Guideline	Item Count
ICML Paper Writing Best Practices[1]	38
ICML 2023 Paper Guidelines[2]	30
NIPS 2024 Reviewer Guidelines[3]	18
ACL Checklist[4]	49
How to Write a Good Research Paper in the Machine Learning Area[5]	6
ACL Ethics Review Questions[6]	21
AAAI Reproducibility Checklist[7]	29
NeurIPS 2021 Paper Checklist Guidelines[8]	46
ICLR 2019 Guidelines[9]	6
Total Count	**243**

A.2.2 Prompts

In this section, we provide prompts designed to annotate reviews. We designed 4 prompts where each corresponds to one of the four combinations of target/aspect and strength/weakness. Specifically, we designed Target-Strength (Figure 9), Aspect-Strength, (Figure 11), Target-Weakness (Figure 10) , and Aspect-Weakness (Figure 12) prompts.

[1] https://icml.cc/Conferences/2022/BestPractices

[2] https://icml.cc/Conferences/2023/PaperGuidelines

[3] https://neurips.cc/Conferences/2024/ReviewerGuidelines

[4] https://aclrollingreview.org/responsibleNLPresearch/

[5] https://www.turing.com/kb/how-to-write-research-paper-in-machine-learning-area

[6] https://2023.eacl.org/ethics/review-questions/

[7] https://aaai.org/conference/aaai/aaai-25/aaai-25-reproducibility-checklist/

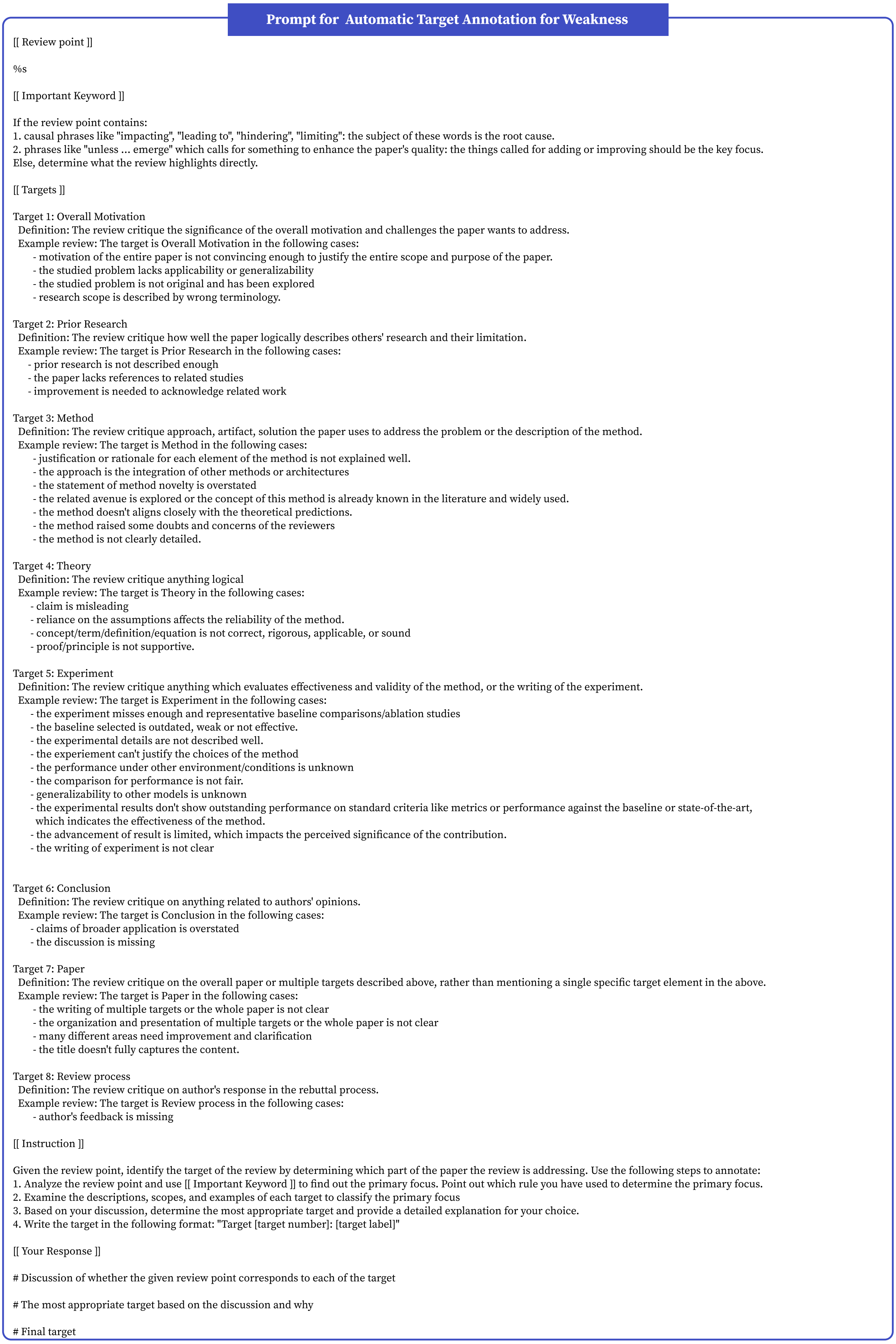
[8] https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

[9] https://iclr.cc/Conferences/2019/Reviewer_Guidelines

Prompt for Automatic Target Annotation for Strength

```
[[ Review point ]]

%s

[[ Important Keyword ]]

If the review point contains:
1. causal phrases like "impacting", "leading to", "demonstrate the merit of": the subject of these words is the root cause.
2. phrases like "is a significant contribution", "making the paper promising" which mark the most important contribution of the paper: the subject modified by these phrases should be the key focus.
Else, determine what the review highlights directly.

[[ Targets ]]

Target 1: Overall Motivation
  Definition: The review praise significance of challenges the paper wants to address
  Example review: The target is Overall Motivation in the following cases:
       - the paper tackles the challenging or important issue/problem
       - the task is practical and innovative

Target 2: Method
  Definition: The review praise the approach, artifact, solution the paper uses to address the problem or the description of the method.
  Example review: The target is Method in the following cases:
       - motivation, intuition, justification or rationale for each element of the method
       - the integration of other methods or architectures is novel
       - the paper identified or addressed an important problem by applying a novel or well-motivated or effective method
       - the method enables the solutions of a challenging problem
       - the method can inspire subsequent research endeavors or has the potential to guide future research
       - the approach exhibits potential for tackling significant problems.
       - the approach opens new avenue
       - the method is rarely explored yet holds significant promise.
       - the method enables exploration into some problems
       - the benefits, implication, generalizability, practical applicability, application of the method
       - the method is clearly detailed.
       - the method aligns closely with the theory
       - the method outperforms the baseline

Target 4: Theory
  Definition: The review praise anything logical.
  Example review: The target is Theory in the following cases:.
      - proof/principle is supportive.
      - theory/concept is novel, impactful, applicable, clear, robust
      - theoretical exploration is valuable

Target 5: Experiment
  Definition: The review praise anything which evaluates effectiveness and validity of the method.
  Example review: The target is Experiment in the following cases:
      - experiments is extensive, comprehensive
      - the experimental results show outstanding performance on standard criteria like metrics or performance against the baseline or state-of-the-art, which indicates the effectiveness of the method.
      - whether the experiment results and their analysis are sound and effective
      - the dataset used in the experiment is novel
      - the experimental results is impactful

Target 6: Conclusion
  Definition: The review praise on anything related to authors' opinions.
  Example review: The target is Conclusion in the following cases:
        - the paper presents promising insights to a important field or domain
        - the author provides insights derived from the experiment results and analysis.
        - the insights are novel, impactful,promising, applicable, appreciated by reviewers, complementing the current understanding, contributing to the community.
        - the authors' interpretation of the results are sound or insightful
        - the paper offers guidelines and suggestions
        - the paper promotes discussions
        - the implication of the results is useful, novel, or insightful
        - the paper identifies key problems in the field

Target 7: Paper
  Definition: The review praise on the overall paper or multiple targets described above, rather than mentioning a single specific target element in the above.
  Example review: The target is Paper in the following cases:
       - the writing of multiple targets or the whole paper is clear, without only saying one target is clear
       - the organization and presentation of multiple targets or the whole paper is clear

Target 8: Review process
  Definition: The review contains praise on author's response, or reviewer's judgement of paper acceptance in the rebuttal process.
  Example review: The target is Review process in the following cases:
        - the authors explain their method clearyly during the rebuttal process
        - the authors actively engaged in the review process
        - the authors' explanation enhanced the paper in the terms of clarty, soundness, impact, completeness, or novelty.
        - all the issues and feedback from preovious reviews were resolved during the review process
        - positive responses and acceptance ratings from reviewers

[[ Instruction ]]

Given the review point, identify the target of the review by determining which part of the paper the review is addressing. Use the following steps to annotate:
1. Analyze the review point and use [[ Important Keyword ]] to find out the primary focus. Point out which rule you have used to determine the primary focus.
2. Examine the descriptions, scopes, and examples of each target to classify the primary focus
3. Based on your discussion, determine the most appropriate target and provide a detailed explanation for your choice.
4. Write the target in the following format: "Target [target number]: [target label]"

[[ Your Response ]]

# Discussion of whether the given review point corresponds to each of the target

# The most appropriate target based on the discussion and why

# Final target
```

Figure 9: Prompt for Automatic Target Annotation for Strength



Prompt for Automatic Target Annotation for Weakness

```
[[ Review point ]]

%s

[[ Important Keyword ]]

If the review point contains:
1. causal phrases like "impacting", "leading to", "hindering", "limiting": the subject of these words is the root cause.
2. phrases like "unless ... emerge" which calls for something to enhance the paper's quality: the things called for adding or improving should be the key focus.
Else, determine what the review highlights directly.

[[ Targets ]]

Target 1: Overall Motivation
  Definition: The review critique the significance of the overall motivation and challenges the paper wants to address.
  Example review: The target is Overall Motivation in the following cases:
       - motivation of the entire paper is not convincing enough to justify the entire scope and purpose of the paper.
       - the studied problem lacks applicability or generalizability
       - the studied problem is not original and has been explored
       - research scope is described by wrong terminology.

Target 2: Prior Research
  Definition: The review critique how well the paper logically describes others' research and their limitation.
  Example review: The target is Prior Research in the following cases:
     - prior research is not described enough
     - the paper lacks references to related studies
     - improvement is needed to acknowledge related work

Target 3: Method
  Definition: The review critique approach, artifact, solution the paper uses to address the problem or the description of the method.
  Example review: The target is Method in the following cases:
       - justification or rationale for each element of the method is not explained well.
       - the approach is the integration of other methods or architectures
       - the statement of method novelty is overstated
       - the related avenue is explored or the concept of this method is already known in the literature and widely used.
       - the method doesn't aligns closely with the theoretical predictions.
       - the method raised some doubts and concerns of the reviewers
       - the method is not clearly detailed.

Target 4: Theory
  Definition: The review critique anything logical
  Example review: The target is Theory in the following cases:
      - claim is misleading
      - reliance on the assumptions affects the reliability of the method.
      - concept/term/definition/equation is not correct, rigorous, applicable, or sound
      - proof/principle is not supportive.

Target 5: Experiment
  Definition: The review critique anything which evaluates effectiveness and validity of the method, or the writing of the experiment.
  Example review: The target is Experiment in the following cases:
      - the experiment misses enough and representative baseline comparisons/ablation studies
      - the baseline selected is outdated, weak or not effective.
      - the experimental details are not described well.
      - the experiement can't justify the choices of the method
      - the performance under other environment/conditions is unknown
      - the comparison for performance is not fair.
      - generalizability to other models is unknown
      - the experimental results don't show outstanding performance on standard criteria like metrics or performance against the baseline or state-of-the-art,
        which indicates the effectiveness of the method.
      - the advancement of result is limited, which impacts the perceived significance of the contribution.
      - the writing of experiment is not clear


Target 6: Conclusion
  Definition: The review critique on anything related to authors' opinions.
  Example review: The target is Conclusion in the following cases:
      - claims of broader application is overstated
      - the discussion is missing

Target 7: Paper
  Definition: The review critique on the overall paper or multiple targets described above, rather than mentioning a single specific target element in the above.
  Example review: The target is Paper in the following cases:
       - the writing of multiple targets or the whole paper is not clear
       - the organization and presentation of multiple targets or the whole paper is not clear
       - many different areas need improvement and clarification
       - the title doesn't fully captures the content.

Target 8: Review process
  Definition: The review critique on author's response in the rebuttal process.
  Example review: The target is Review process in the following cases:
       - author's feedback is missing

[[ Instruction ]]

Given the review point, identify the target of the review by determining which part of the paper the review is addressing. Use the following steps to annotate:
1. Analyze the review point and use [[ Important Keyword ]] to find out the primary focus. Point out which rule you have used to determine the primary focus.
2. Examine the descriptions, scopes, and examples of each target to classify the primary focus
3. Based on your discussion, determine the most appropriate target and provide a detailed explanation for your choice.
4. Write the target in the following format: "Target [target number]: [target label]"

[[ Your Response ]]

# Discussion of whether the given review point corresponds to each of the target

# The most appropriate target based on the discussion and why

# Final target
```

Figure 10: Prompt for Automatic Target Annotation for Weakness

Prompt for Automatic Aspect Annotation for Strength

```
[[ Review point ]]
%s

[[ Aspects ]]

Aspect 1: Impact
  Definition: The review explicitly praises how paper influences future research, researchers, or practitioners
  Example review: The aspect is Impact in the following cases:
    - The paper opens new important avenue or suggests novel perspectives that has not been explored
    - The paper makes a breakthrough in the field
    - The method has practical utility
    - The method is generally applicable in various use cases
    - The theory offers generalizable insights
    - The paper tackles one of the most challenging problem in the field


Aspect 2: Novelty
  Definition: The review explicitly praises the originality of the contributions, compared to existing knowledge.
  Example review: The aspect is Novelty in the following cases:
    - The author addresses overlooked, but important problems
    - The method is new and useful, compared to existing methods
    - The theory offers new insights, that have not been previously known
    - The experiment setting is unconventional, offering novel insights

Aspect 3: Communication Clarity
  Definition: The review explicitly praises how clearly the author communicates ideas
  Example review: The aspect is Communication Clarity in the following cases:
    - The paper is clear and well-structured
    - The method is clearly described
    - The theory is easy to understand

Aspect 4: Validity
  Definition: The review explicitly praises effectiveness or soundness of research
  Example review: The aspect is Validity in the following cases:
    - The paper introduces effective methods
    - The paper introduces theories with proof
    - The problem statement is sound
    - The experiment clearly shows that the method outperforms existing methods
    - The methodology is sound and clear
    - The experiment is comprehensively done
    - The author claims are supported or justified well
    - The theory is clear and convincing

Aspect 5: Not-specific
  Definition: The review generally praises multiple aspects, rather than emphasizing a single specific aspect in the above.
  Exaple review: The aspect is Not-specific in the following cases:
    - The paper is high-quality in terms of its validity, novelty, and impact
    - The paper presents novel methods with valid methdoology
    - The paper presents convincing arguments with practical impact

Aspect 6: Irrelevant
  Definition: The review does not pertain to the evaluation of the paper's content, contributions, or quality, but rather discuss a events in the rebuttal process

[[ Instruction ]]

Given the review point, critically identify the aspect of the review by determining which characteristic of the paper the review is addressing. Use the following steps to annotate:

1. For each potential aspect, discuss whether the review directly and explicitly corresponds to the aspect. Highlight why the review point supports or contradicts the aspect.
2. Based on your discussion, discuss the most appropriate aspect, focusing on the main subject of the praise.
3. Write the aspect in the following format: ""Aspect [aspect number]: [aspect label]""

[[ Your Response ]]

# Discussion of whether the review point corresponds to each of the aspect
## Aspect 1: Impact
- (a single paragraph of the discussion)

## Aspect 2: Novelty
- (a single paragraph of the discussion)

...

# The most appropriate aspect based on the discussion on the review point and why

# Final aspect
```

Figure 11: Prompt for Automatic Aspect Annotation for Strength

Prompt for Automatic Aspect Annotation for Weakness

```
[[ Review point ]]
%s

[[ Aspects ]]

Aspect 1: Validity
  Definition: The review explicitly critiques completeness, soundness, or validity of research
  Example review: The aspect is Validity in the following cases:
    - The problem statement lacks definition
    - The prior work has not been comprehensively surveyed
    - The method lacks justification
    - The experiment does not show the effectiveness of the method, compared to existing methods
    - The scope of experiment is too narrow, limiting its applicability
    - The claim lacks justifications or sufficient evidences to be supported
    - The assumptions are not realistic

Aspect 2: Communication Clarity
  Definition: The review explicitly critiques how clearly the author communicates ideas
  Example review: The aspect is Communication Clarity in the following cases:
    - The paper does not provide clear explanations about rationale
    - The paper uses unclear terminology
    - The method description is ambiguous or lacks details
    - The description of theory is not clear
    - The paper is difficult to understand
    - Some of the claims are misleading
    - Lack of comprehensive examples make it difficult to understand the paper

Aspect 3: Novelty
  Definition: The review explicitly critiques the originality of the contributions, compared to existing knowledge.
  Example review: The aspect is Novelty in the following cases:
    - The method is a straightforward extension of prior work
    - The theory is not new and useful, compared to existing theories
    - The experiments and insights are already known in prior work

Aspect 4: Impact
  Definition: The review explicitly critiques how paper influences future research, researchers, or practitioners
  Example review: The aspect is Impact in the following cases:
    - The method is not applicable nor generalizable
    - The method is not easily extended to real-world scenarios
    - The insights are not practically useful


Aspect 5: Not-specific
  Definition: The review generally critiques multiple aspects, rather than emphasizing a single specific aspect in the above.
  Exaple review: The aspect is Not-specific in the following cases:
    - Reviewers have a consensus for rejection, criticizing the validity and clarity of the proposed methods
    - The paper needs significant revisions, including justifying their methods, better positioning for novelty, and clearly outlining their implications
    - The paper needs to clarity the study setup and enhance the readibility in sections

Aspect 6: Irrelevant
  Definition: The review does not pertain to the evaluation of the paper's content, contributions, or quality, but rather discuss a events in the rebuttal process

[[ Instruction ]]

Given the review point, critically identify the aspect of the review by determining which characteristic of the paper the review is addressing. Use the following steps to annotate:

1. For each potential aspect, discuss whether the review directly and explicitly corresponds to the aspect. Highlight why the review point supports or contradicts the aspect.
2. Based on your discussion, discuss the most appropriate aspect, focusing on the main subject of the critique.
3. Write the aspect in the following format: ""Aspect [aspect number]: [aspect label]""

[[ Your Response ]]

# Discussion of whether the review point corresponds to each of the aspect
## Aspect 1: Validity
- (a single paragraph of the discussion)

## Aspect 2: Communication Clarity
- (a single paragraph of the discussion)

...

# The most appropriate aspect based on the discussion on the review point and why

# Final aspect
```

Figure 12: Prompt for Automatic Aspect Annotation for Weakness

A.2.3 Annotation Comparison

We present a comparison between LLM and human annotations for both target and aspect. Figures 13 and Figure 14 illustrate the discrepancies. Areas of alignment between LLM and human annotations are shown in green, while red highlights regions with significant discrepancies.

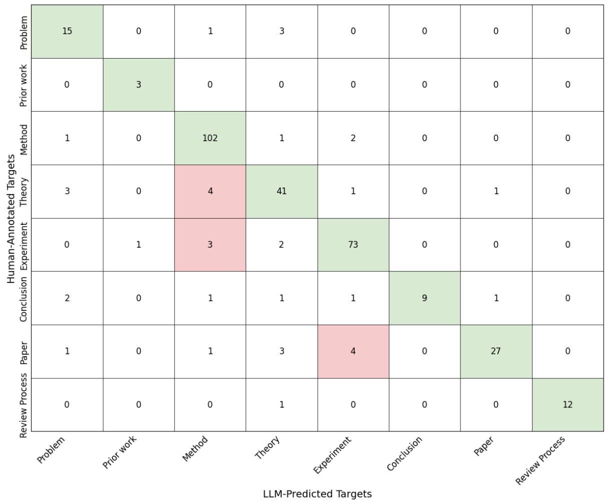


Figure 13: LLM vs. human target annotation

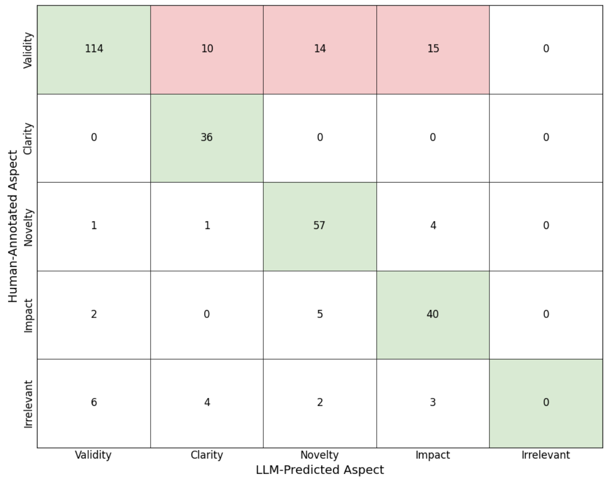
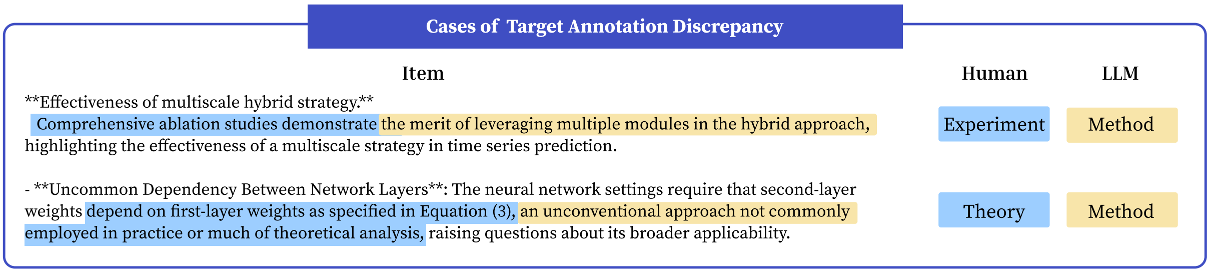


Figure 14: LLM vs. human aspect annotation

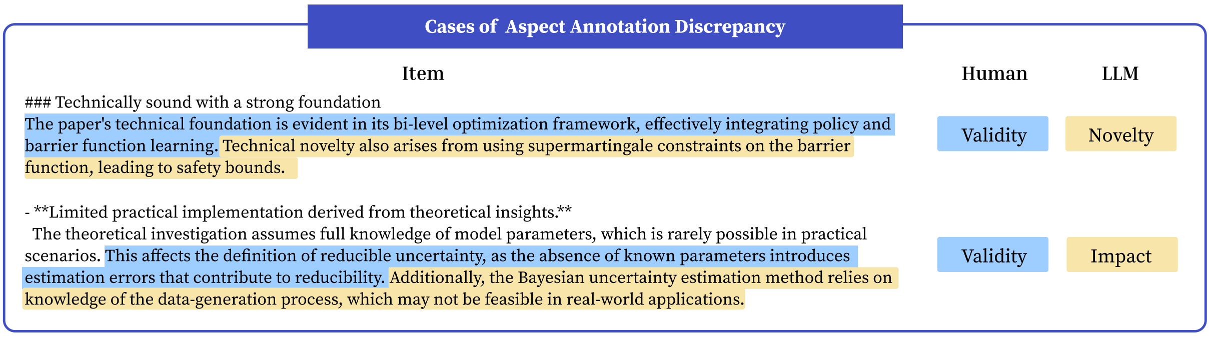
While LLM annotations differ from human annotations in some cases, certain discrepancies remain reasonable. Figure 15 and Figure 16 illustrate examples of such reasonable discrepancies.



Cases of Target Annotation Discrepancy

Item	Human	LLM
Effectiveness of multiscale hybrid strategy. Comprehensive ablation studies demonstrate the merit of leveraging multiple modules in the hybrid approach, highlighting the effectiveness of a multiscale strategy in time series prediction.	Experiment	Method
- **Uncommon Dependency Between Network Layers**: The neural network settings require that second-layer weights depend on first-layer weights as specified in Equation (3), an unconventional approach not commonly employed in practice or much of theoretical analysis, raising questions about its broader applicability.	Theory	Method

Figure 15: Cases of Target Annotation Discrepancy



Cases of Aspect Annotation Discrepancy

Item	Human	LLM
### Technically sound with a strong foundation The paper's technical foundation is evident in its bi-level optimization framework, effectively integrating policy and barrier function learning. Technical novelty also arises from using supermartingale constraints on the barrier function, leading to safety bounds.	Validity	Novelty
- **Limited practical implementation derived from theoretical insights.** The theoretical investigation assumes full knowledge of model parameters, which is rarely possible in practical scenarios. This affects the definition of reducible uncertainty, as the absence of known parameters introduces estimation errors that contribute to reducibility. Additionally, the Bayesian uncertainty estimation method relies on knowledge of the data-generation process, which may not be feasible in real-world applications.	Validity	Impact

Figure 16: Cases of Aspect Annotation Discrepancy

A.2.4 Results

The following tables present a comprehensive performance comparison of models across different metrics and evaluation targets, including both strengths and weaknesses (Table 7), as well as separate analyses focusing on strengths (Table 8) and weaknesses (Table 9). Additionally, we provide a similar comparison across metrics and broader aspects, including both strengths and weaknesses (Table 10), strengths alone (Table 11), and weaknesses alone (Table 12).

Table 7: Performance Comparison of Models Across Metrics and Targets (Including both Strengths and Weaknesses)

Target	Problem	Prior Research	Method	Theory	Experiment	Conclusion	Paper
F1 (gpt-4o-mini)	0.268	0.076	0.737	0.427	0.680	0.103	0.227
F1 (gpt-4o)	0.292	0.052	0.741	0.448	0.673	0.089	0.247
F1 (o1-mini)	0.275	0.054	**0.764**	0.472	**0.684**	**0.175**	**0.253**
F1 (o1)	0.274	0.044	0.754	**0.489**	0.673	0.133	0.091
F1 (llama-70B)	0.269	0.049	0.711	0.410	0.659	0.172	0.158
F1 (llama-405B)	0.158	0.031	0.690	0.427	0.662	0.167	0.134
F1 (deepseek-r1)	**0.297**	**0.081**	0.729	0.473	0.682	0.164	0.152
F1 (deepseek-v3)	0.241	0.051	0.725	0.405	0.680	0.110	0.092
Prec (gpt-4o-mini)	0.317	0.134	0.647	0.317	0.549	0.063	0.241
Prec (gpt-4o)	0.298	0.109	0.634	0.334	0.547	0.057	0.251
Prec (o1-mini)	0.315	0.130	0.639	0.342	0.549	0.107	0.274
Prec (o1)	0.279	0.064	**0.648**	**0.381**	0.549	0.111	0.245
Prec (llama-70B)	**0.339**	**0.143**	0.653	0.295	0.548	0.105	0.289
Prec (llama-405B)	0.324	0.071	0.647	0.310	**0.558**	0.115	0.233
Prec (deepseek-r1)	0.321	0.099	0.639	0.327	0.549	**0.135**	**0.301**
Prec (deepseek-v3)	0.288	0.100	0.645	0.280	0.547	0.076	0.249
Rec (gpt-4o-mini)	0.233	0.053	0.870	0.691	0.983	0.274	0.232
Rec (gpt-4o)	0.297	0.034	0.899	0.723	0.965	0.202	**0.270**
Rec (o1-mini)	0.266	0.034	**0.952**	0.834	**0.994**	**0.536**	0.249
Rec (o1)	**0.353**	0.034	0.905	0.736	0.963	0.167	0.056
Rec (llama-70B)	0.246	0.030	0.803	0.720	0.919	0.476	0.146
Rec (llama-405B)	0.108	0.020	0.774	0.694	0.894	0.300	0.095
Rec (deepseek-r1)	0.299	**0.069**	0.859	**0.865**	0.983	0.357	0.102
Rec (deepseek-v3)	0.210	0.035	0.844	0.755	0.981	0.238	0.058

Table 8: Performance Comparison of Models Across Metrics and Targets (Strengths)

Target	Problem	Prior Research	Method	Theory	Experiment	Conclusion	Paper
F1 (gpt-4o-mini)	0.283	0.000	**0.760**	0.424	0.511	0.118	0.232
F1 (gpt-4o)	0.329	0.000	0.756	0.446	**0.517**	0.143	0.119
F1 (o1-mini)	0.345	0.000	0.753	0.411	0.511	0.300	**0.233**
F1 (o1)	0.384	0.000	0.749	**0.470**	0.512	0.267	0.061
F1 (llama-70B)	0.245	0.000	0.750	0.420	0.516	0.242	0.198
F1 (llama-405B)	0.160	0.000	0.755	0.455	0.516	**0.333**	0.079
F1 (deepseek-r1)	**0.396**	0.000	0.749	0.436	0.513	0.174	0.135
F1 (deepseek-v3)	0.331	0.000	0.755	0.423	0.509	0.114	0.086
Prec (gpt-4o-mini)	0.315	0.000	0.622	0.286	0.343	0.071	0.198
Prec (gpt-4o)	0.295	0.000	0.616	0.299	0.350	0.091	0.182
Prec (o1-mini)	0.314	0.000	0.611	0.264	0.343	0.176	0.203
Prec (o1)	0.285	0.000	**0.624**	**0.322**	0.346	0.222	0.172
Prec (llama-70B)	0.404	0.000	0.620	0.275	0.352	0.148	0.178
Prec (llama-405B)	**0.419**	0.000	0.620	0.319	**0.358**	**0.231**	0.163
Prec (deepseek-r1)	0.355	0.000	0.617	0.289	0.347	0.103	**0.279**
Prec (deepseek-v3)	0.364	0.000	0.620	0.276	0.344	0.069	0.154
Rec (gpt-4o-mini)	0.258	0.000	0.975	0.819	**0.996**	0.333	**0.281**
Rec (gpt-4o)	0.371	0.000	0.978	0.872	0.991	0.333	0.089
Rec (o1-mini)	0.382	0.000	**0.980**	**0.935**	**0.996**	**1.000**	0.274
Rec (o1)	**0.588**	0.000	0.936	0.872	0.987	0.333	0.037
Rec (llama-70B)	0.176	0.000	0.948	0.894	0.969	0.667	0.224
Rec (llama-405B)	0.099	0.000	0.965	0.796	0.921	0.600	0.052
Rec (deepseek-r1)	0.447	0.000	0.953	0.883	0.983	0.571	0.089
Rec (deepseek-v3)	0.303	0.000	0.963	0.904	0.982	0.333	0.059

Table 9: Performance Comparison of Models Across Metrics and Targets (Weaknesses)

Target	Problem	Prior Research	Method	Theory	Experiment	Conclusion	Paper
F1 (gpt-4o-mini)	0.253	0.153	0.715	0.430	0.849	0.088	0.222
F1 (gpt-4o)	0.256	0.104	0.726	0.449	0.830	0.036	**0.375**
F1 (o1-mini)	0.204	0.108	**0.774**	0.534	**0.857**	0.050	0.272
F1 (o1)	0.164	0.089	0.760	0.508	0.835	0.000	0.120
F1 (llama-70B)	**0.294**	0.098	0.672	0.400	0.802	0.103	0.118
F1 (llama-405B)	0.155	0.062	0.625	0.399	0.809	0.000	0.190
F1 (deepseek-r1)	0.198	**0.163**	0.709	**0.510**	0.852	**0.154**	0.169
F1 (deepseek-v3)	0.151	0.103	0.696	0.387	0.850	0.105	0.099
Prec (gpt-4o-mini)	**0.320**	0.268	0.672	0.347	0.755	0.056	0.283
Prec (gpt-4o)	0.301	0.219	0.651	0.369	0.743	0.024	0.321
Prec (o1-mini)	0.315	0.259	0.666	0.420	0.754	0.038	0.345
Prec (o1)	0.273	0.127	0.672	**0.440**	0.752	0.000	0.317
Prec (llama-70B)	0.274	**0.286**	**0.687**	0.315	0.744	0.062	**0.400**
Prec (llama-405B)	0.228	0.143	0.673	0.300	**0.758**	0.000	0.304
Prec (deepseek-r1)	0.287	0.197	0.661	0.365	0.750	**0.167**	0.323
Prec (deepseek-v3)	0.212	0.200	0.669	0.284	0.750	0.083	0.345
Rec (gpt-4o-mini)	0.209	0.107	0.764	0.563	0.970	**0.214**	0.183
Rec (gpt-4o)	0.222	0.068	0.821	0.574	0.939	0.071	**0.451**
Rec (o1-mini)	0.151	0.068	**0.924**	0.732	**0.992**	0.071	0.224
Rec (o1)	0.118	0.068	0.874	0.600	0.939	0.000	0.074
Rec (llama-70B)	**0.316**	0.059	0.658	0.547	0.869	0.286	0.069
Rec (llama-405B)	0.118	0.040	0.583	0.593	0.867	0.000	0.138
Rec (deepseek-r1)	0.151	**0.139**	0.764	**0.847**	0.984	0.143	0.115
Rec (deepseek-v3)	0.118	0.069	0.725	0.605	0.980	0.143	0.057

Table 10: Performance Comparison of Models Across Metrics and Aspects (Including both Strengths and Weaknesses)

Aspect	Novelty	Impact	Validity	Clarity
F1 (gpt-4o-mini)	0.334	0.390	**0.775**	0.396
F1 (gpt-4o)	0.378	**0.428**	0.769	0.365
F1 (o1-mini)	0.386	0.427	0.773	0.395
F1 (o1)	**0.404**	0.399	0.772	**0.401**
F1 (llama-70B)	0.334	0.322	0.769	0.327
F1 (llama-405B)	0.337	0.318	0.772	0.278
F1 (deepseek-r1)	0.387	0.414	**0.775**	0.266
F1 (deepseek-v3)	0.346	0.422	0.768	0.187
Prec (gpt-4o-mini)	0.367	0.291	**0.671**	0.317
Prec (gpt-4o)	0.474	0.313	0.668	0.298
Prec (o1-mini)	0.528	0.300	0.668	0.311
Prec (o1)	0.589	0.305	0.669	0.334
Prec (llama-70B)	**0.665**	**0.318**	0.667	0.337
Prec (llama-405B)	0.587	0.302	**0.671**	0.332
Prec (deepseek-r1)	0.535	0.308	0.670	**0.339**
Prec (deepseek-v3)	0.504	0.306	0.664	0.309
Rec (gpt-4o-mini)	0.460	0.600	**0.990**	**0.549**
Rec (gpt-4o)	0.506	0.689	0.975	0.485
Rec (o1-mini)	**0.507**	**0.758**	**0.990**	0.548
Rec (o1)	0.435	0.579	0.981	0.511
Rec (llama-70B)	0.450	0.371	0.981	0.346
Rec (llama-405B)	0.478	0.352	0.978	0.241
Rec (deepseek-r1)	0.502	0.632	0.988	0.219
Rec (deepseek-v3)	0.478	0.683	0.982	0.134

Table 11: Performance Comparison of Models Across Metrics and Aspects (Strengths)

Aspect	Novelty	Impact	Validity	Clarity
F1 (gpt-4o-mini)	0.643	0.474	**0.599**	0.309
F1 (gpt-4o)	0.654	0.520	0.593	0.202
F1 (o1-mini)	0.656	**0.556**	0.592	0.299
F1 (o1)	0.626	0.530	0.596	**0.342**
F1 (llama-70B)	0.636	0.411	0.593	0.292
F1 (llama-405B)	**0.660**	0.345	0.596	0.157
F1 (deepseek-r1)	0.655	0.536	0.598	0.170
F1 (deepseek-v3)	**0.660**	0.547	0.585	0.122
Prec (gpt-4o-mini)	0.498	0.368	**0.431**	0.222
Prec (gpt-4o)	0.498	0.398	0.428	0.190
Prec (o1-mini)	0.501	0.403	0.424	0.224
Prec (o1)	**0.530**	0.412	0.430	**0.261**
Prec (llama-70B)	0.497	**0.467**	0.426	0.236
Prec (llama-405B)	0.506	0.368	**0.431**	0.215
Prec (deepseek-r1)	0.503	0.400	**0.431**	0.224
Prec (deepseek-v3)	0.509	0.403	0.419	0.207
Rec (gpt-4o-mini)	0.907	0.667	**0.986**	**0.511**
Rec (gpt-4o)	**0.955**	0.749	0.965	0.216
Rec (o1-mini)	0.949	**0.897**	0.979	0.449
Rec (o1)	0.763	0.744	0.969	0.496
Rec (llama-70B)	0.883	0.366	0.976	0.384
Rec (llama-405B)	0.949	0.324	0.969	0.123
Rec (deepseek-r1)	0.937	0.809	0.976	0.137
Rec (deepseek-v3)	0.940	0.851	0.965	0.086

Table 12: Performance Comparison of Models Across Metrics and Aspects (Weaknesses)

Aspect	Novelty	Impact	Validity	Clarity
F1 (gpt-4o-mini)	0.024	0.306	0.951	0.484
F1 (gpt-4o)	0.103	**0.335**	0.945	**0.528**
F1 (o1-mini)	0.116	0.299	**0.954**	0.492
F1 (o1)	**0.182**	0.268	0.949	0.459
F1 (llama-70B)	0.032	0.233	0.945	0.362
F1 (llama-405B)	0.013	0.291	0.947	0.399
F1 (deepseek-r1)	0.120	0.292	0.952	0.362
F1 (deepseek-v3)	0.031	0.297	0.951	0.253
Prec (gpt-4o-mini)	0.235	0.214	**0.912**	0.411
Prec (gpt-4o)	0.450	0.228	0.907	0.406
Prec (o1-mini)	0.556	0.197	0.911	0.397
Prec (o1)	0.647	0.198	0.908	0.406
Prec (llama-70B)	**0.833**	0.169	0.907	0.438
Prec (llama-405B)	0.667	**0.236**	0.911	0.450
Prec (deepseek-r1)	0.568	0.215	0.908	**0.454**
Prec (deepseek-v3)	0.500	0.209	0.908	0.410
Rec (gpt-4o-mini)	0.013	0.533	0.994	0.587
Rec (gpt-4o)	0.058	**0.630**	0.985	**0.754**
Rec (o1-mini)	0.065	0.619	**1.000**	0.646
Rec (o1)	**0.106**	0.415	0.994	0.527
Rec (llama-70B)	0.016	0.376	0.987	0.308
Rec (llama-405B)	0.006	0.381	0.987	0.359
Rec (deepseek-r1)	0.067	0.455	**1.000**	0.302
Rec (deepseek-v3)	0.016	0.515	0.998	0.183