Mind the Blind Spots:
A Focus-Level Evaluation Framework for LLM Reviews

Anonymous ACL submission

Abstract

Peer review underpins scientific progress, but it is increasingly strained by reviewer shortages and growing workloads. Large Language Models (LLMs) can automatically draft reviews now, but determining whether LLM-generated reviews are trustworthy requires systematic evaluation. Researchers have evaluated LLM reviews at either surface-level (e.g., BLEU and ROUGE) or content-level (e.g., specificity and factual accuracy). Yet it remains uncertain whether LLM-generated reviews attend to the same critical facets that human experts weigh—the strengths and weaknesses that ultimately drive an accept-or-reject decision. We introduce a focus-level evaluation framework that operationalizes the focus as a normalized distribution of attention across predefined facets in paper reviews. Based on the framework, we developed an automatic focus-level evaluation pipeline based on two sets of facets: target (e.g., problem, method, and experiment) and aspect (e.g., validity, clarity, and novelty), leveraging 676 paper reviews[1] from OpenReview that consists of 3,657 strengths and weaknesses identified from human experts. The comparison of focus distributions between LLMs and human experts showed that the off-the-shelf LLMs consistently have a more biased focus towards examining technical validity while significantly overlooking novelty assessment when criticizing papers.

1 Introduction

Reviewing academic papers lies at the heart of scientific advancement, but it requires 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). Large Language Models (LLMs) hold the potential to assist the peer review process by automatically reviewing papers (Hosseini and Horbach, 2023; Robertson, 2023), but can we trust LLM-generated reviews? Evaluating the quality of reviews is inherently complex due to their multi-dimensional nature. Researchers have employed various metrics for the evaluation such as surface-level (e.g. linguistic similarity to human reviews), content-level (e.g., relevance, specificity, and factual accuracy), and decision-level (e.g., accept/reject classification accuracy) metrics (Ramachandran et al., 2017; Du et al., 2024; Liang et al., 2024; Zhou et al., 2024).

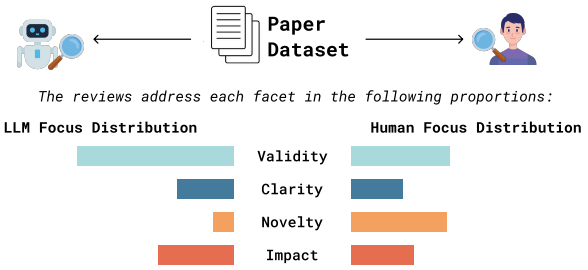

Figure 1: We introduce a focus-level evaluation framework for assessing LLM reviews, which computes focus distributions and compares them against human reviews based on predefined facets. The focus-level evaluation offers actionable insights into how to improve LLMs' paper review capability and how to most effectively leverage LLM reviews in the peer review process.

However, existing evaluations fail to assess whether LLM reviews comprehensively address critical dimensions of papers. Evaluating the *focus* of reviews is crucial because reviews with poor focus can negatively impact reviewers, even if they are accurate, relevant, and specific. For example, reviews that overly concentrate on methodological details while completely neglecting the novelty aspect of the proposed method could fail to suggest meaningful feedback, diverging from how expert reviewers assess the submission. It could also mislead junior reviewers by promoting incomplete

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

perspectives and reinforce shallow paper review practices. Despite such importance, few attempts have been made to systematically evaluate whether the focus of LLM reviews aligns with that of expert reviews. Conducting the **focus-level evaluation** of LLM reviews is useful to reveal the blind spots of LLM reviews along with their central focus, offering important insights into how human reviewers can most effectively leverage LLM reviews in the peer review process. Moreover, it provides a concrete foundation for guiding LLM training toward more balanced and expert-aligned review behavior.

We introduce a framework for focus-level evaluation of LLM reviews, which systematically analyzes where the reviews direct their praise and criticism based on facets considered important in peer review (Figure 1). Given an LLM, the framework computes a **focus distribution**, a normalized distribution of how frequently review points (e.g., a list of strengths and weaknesses) address predefined facets (e.g., problem, method, and experiments) by leveraging a paper review dataset. The focus distribution can be computed by an automatic annotator that assigns a facet for each review point, enabling a fully automatic evaluation. The interpretable nature of the focus distribution provides actionable insights by clearly revealing which facets LLMs tend to emphasize or overlook in comparison to human experts.

To apply this framework for analyzing LLM-generated reviews in the context of AI conferences, we implemented a focus-level evaluation pipeline (Figure 2). We identified the facets that constitute review focus, by surveying 9 paper submission guidelines from AI conferences and prior literature on review analysis (Chakraborty et al., 2020; Ghosal et al., 2022; Yuan et al., 2022). We define two sets of facets: target (*what* review points praise and critique such as problem, method, and experiment) and aspect (*which criteria* is being evaluated such as validity, clarity, and novelty), which are key elements in analyzing paper reviews (Ghosal et al., 2022; Lu et al., 2025). We identified 7 facets for the target and 5 facets for the aspect (Table 1). Next, we developed an automatic annotator for computing the focus distributions based on the target and aspect, which assigns a target and aspect label for a strength and weakness point in a review. The annotator showed substantial agreement with human annotators, achieving IRR (Cohen's kappa (Cohen, 1960)) of 0.81 for target and 0.79 for aspect.

As a benchmark dataset for our focus-level evaluation pipeline, we constructed a dataset of 676 papers and their review data from OpenReview for ICLR conferences spanning 2021 to 2024. Then we computed and compared the focus distributions of human and LLM reviews using the evaluation pipeline (Figure 4), and we also measured text similarities between the reviews. Specifically, we evaluated 8 LLMs (4 GPT, 2 Llama, and 2 DeepSeek family) to analyze their review focus. We also evaluated MARG (D'Arcy et al., 2024) as a novel review generation technique and a fine-tuned `gpt-4o` using our dataset. The results showed that:

- LLMs struggle to identify key targets and aspects in their reviews. Even the top-performing model reached an F1 score of 0.373 when matching human reviewers on the targets and aspects in each review point.
- LLMs' review focus was biased towards examining technical validity, *consistently overlooking novelty assessment* in weaknesses – a critical limitation in paper review.
- The fine-tuned model produced focus distributions most closely aligned with that of humans, compared to models using prompting alone.
- The models demonstrated strengths in distinct areas. While the fine-tuned model produced the closest focus distributions, `Llama-405B` achieved the highest text similarity. It highlights the importance of holistic evaluation to capture the diverse aspects of review quality.

We release a dataset comprising 676 papers, expert reviews, 3,657 strengths and weaknesses identified from the expert reviews with automatically annotated targets and aspects, LLM-generated reviews from 8 LLMs, and a total of 43,042 strengths and weaknesses extracted from the LLMs, each annotated with corresponding targets and aspects.

2 A Framework for Focus-Level Evaluation of LLM Reviews

We propose a *focus-level evaluation* framework to systematically analyze what aspects LLMs emphasize or overlook when reviewing scientific papers. To enable interpretable and automated assessments of LLM behavior in reviewing, we aim to reveal the distribution of attention an LLM allocates to different review facets when identifying strengths and weaknesses in submissions. Specifically, we

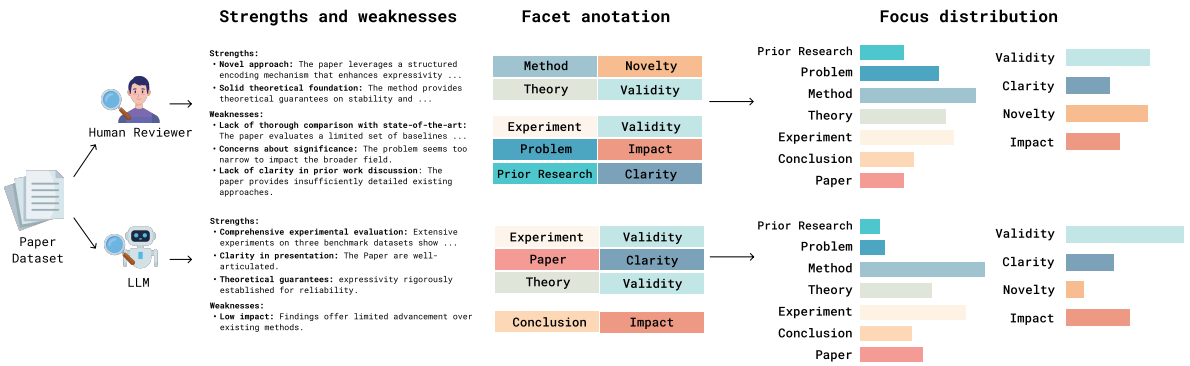

Figure 2: The overall process of automated focus-level evaluation. We first extracted strengths and weaknesses from review data on the OpenReview platform as the expert reviews. To identify key strengths and weaknesses influencing the final acceptance, we extracted them from the meta-review and augmented details from individual reviewer comments. Each strength and weakness was then annotated with a target and aspect by our automatic annotator. Finally, we computed the focus distributions by normalizing the frequency of annotated targets and aspects, and compare this distribution with that of LLM reviews.

define a focus of the review to be compared for focus-level evaluation as follows:

Let (i) L be an LLM, (ii) $A = \{a_1, a_2, ..., a_N\}$ be a list of facets where each facet denotes a distinct criteria (e.g., problem, method, and experiment), and (iii) $P = \{p_1, p_2, ..., p_M\}$ be a corpus of paper submissions. The focus-level evaluation $E(L, A, P)$ produces two focus distributions F^+ and F^- where F^+ denotes the distribution when identifying strengths of the submissions and F^- for weaknesses. The focus distribution $F = (f_1, f_2, ..., f_N)$ can be represented as a normalized vector where f_i denotes the relative frequency of review points (i.e., strengths for F^+ and weaknesses for F^-) that discuss the facet a_i, when L generates reviews for paper submissions in P.

To assess LLM behavior, our framework compares focus distributions with those from human expert reviews. Researchers can specify the set of facets A and the paper corpus P based on the goals of their analysis, allowing flexible and targeted focus-evaluation.

Based on this framework, we implement an automatic focus-level evaluation pipeline to understand LLM's behavior in reviewing AI papers. Figure 2 illustrates the process of our focus-level evaluation pipeline. Our approach consists of three steps. (i) Collect an expert review dataset from ICLR conferences and extract strengths and weaknesses of the submissions for computing focus distributions of human experts (Section 3), (ii) Define facets based on paper submission guidelines of AI conferences and build an automatic annotator based on the facets (Section 4), and (iii) Compute and analyze the focus distributions of LLMs and human experts in reviewing AI papers (Section 5).

3 Constructing Expert Review Dataset

The focus-level evaluation framework requires a corpus of paper submissions P. We collected the review data from OpenReview platform and extracted the strengths and weaknesses of papers for computing focus distributions of human experts.

3.1 Collecting Review Data

We used real-world review data covering ICLR 2021-2024 from the OpenReview platform[2], where human experts evaluated submissions for a top-tier AI conference. Using the OpenReview API[3] and the list of submissions from public GitHub repositories[4], we initially collected 18,407 submissions with their review data.

3.2 Extracting 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 reviews from several reviewers. To address the challenge, our approach is to use a meta-review, a final review from a qualified expert that summarizes reviews and highlights important strengths and weaknesses for supporting the final

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

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

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

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).

4 Developing an Automatic Focus-level Evaluation Method

To enable a fully automated evaluation using the proposed focus-level evaluation framework, we first define a set of facets and then develop an automatic annotator. We then compute focus distributions based on the annotated facets to analyze how LLMs and human reviewers differ in their focus of reviewing.

4.1 Defining Facets from Guidelines

To build an initial set of facets, 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*). To ensure comprehensive coverage of facets, we also reviewed literature that analyzes 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). The definition of each target and aspect facet is available in Appendix A.2.2.

Target	Aspect
Problem	Impact
Prior Research	Novelty
Method	Clarity
Theory	Validity
Experiment	Not-specific
Conclusion	
Paper	

Table 1: Our research focuses on two sets of facets: target and aspect. Detailed definitions of the facets are available in Appendix A.2.2.

4.2 Building Automatic Annotators

Based on the identified facets, 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 — one author is experienced in qualitative research in HCI and the other author has prior publications in the field of AI/NLP — synchronously decided each label 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 ... Technical novelty also arises from using supermartingale constraints ...*" 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 wrapped in "**".

Model	Target	Aspect
`gpt-4o-mini`	0.69	0.71
`gpt-4o`	0.83	0.75
`o3-mini`	0.81	0.79

Table 2: Inter-Rater Reliability (Cohen's kappa (Cohen, 1960)) between annotations of authors and LLMs.

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.2.3. Table 2 shows the Inter-Rater Reliability (IRR, Cohen's kappa (Cohen, 1960)) between human and LLM annotations for three language models. Annotation using `o3-mini` achieved the IRR scores of 0.81 for targets and 0.79 for aspects, indicating substantial agreement (Cohen, 1960). Given the high IRR and its relatively low computational cost compared to other two models, we used `o3-mini` for the automatic annotation of both target and aspect in the main evaluation. Moreover, an examination of the confusion matrix (Appendix A.2.4) 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.

4.3 Computing Focus Distributions

Building on the defined facets and the automatic annotation method, we assign a target and aspect

Model	Focus similarity				Text similarity		
	KL Divergence	Overall F1	Strength F1	Weakness F1	ROUGE-L	BERTScore	BLEU-4
gpt-4o-mini	0.081	0.344	0.335	0.353	0.197	0.883	0.076
gpt-4o	0.082	0.348	0.342	0.354	0.202	**0.885**	0.079
o1-mini	0.090	0.359	0.331	0.385	0.179	0.878	0.059
o1	0.097	0.355	0.318	0.388	0.170	0.869	0.032
DeepSeek-R1	0.120	**0.373**	0.341	**0.400**	0.156	0.874	0.045
Llama-70B	0.136	0.339	0.338	0.341	0.215	0.882	0.076
Llama-405B	0.145	0.349	**0.349**	0.350	**0.218**	0.884	**0.089**
DeepSeek-V3	0.151	0.350	0.330	0.368	0.199	0.880	0.069
gpt-4o (FT)	**0.022**	0.306	0.280	0.322	0.194	0.882	0.081
MARG	0.113	0.346	–	0.346	0.160	0.854	0.011

Table 3: Overall performance by comparing expert reviews and LLM reviews. For focus similarity, we computed an average of the KL divergences of four focus distributions (strength/target, weakness/target, strength/aspect, and weakness/aspect) between LLM and expert reviews. The overall, strength, and weakness F1 scores were computed by comparing the (target, aspect) set between expert and LLM reviews. The text similariy metrics were computed between LLM reviews and expert reviews. The results highlight different areas of excellence across models (gpt-4o (FT): the highest focus distribution similarity, DeepSeek-R1: the best agreement on (target, aspect) labels, Llama-405B: the highest text similarity score.)

label to each strength and weakness point, using the automatic annotator. We then compute the normalized frequency of these labels to derive focus distributions of targets and aspects, respectively. Separate distributions are calculated for strengths and weaknesses, resulting in four distinct focus distributions. These focus distributions illustrate how LLMs and human reviewers allocate their attention across the different facets of a paper.

5 Evaluation

5.1 Setup

Data. 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.

For *accepted* papers, we manually collected the submitted versions of a small sample (40 papers), which has the timestamp near the ICLR deadline in the version history in arXiv. See Appendix A.5 for the focus distribution results.

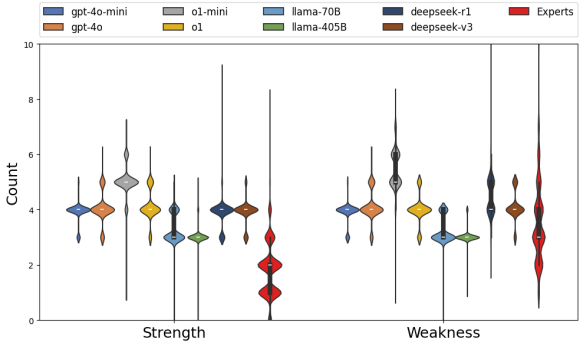

Figure 3: Distribution of strengths and weaknesses. Unlike human experts, LLMs reported a consistent count regardless of paper contents. o1-mini identified the most, while Llama models identified the fewest points.

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, o3-mini, o1)[5], two Llama models (Llama 3.1-{70B, 405B}), and two DeepSeek models (DeepSeek-{V3, R1}). We also evaluated MARG (D'Arcy et al., 2024) and a fine-tuned gpt-4o (see Appendix A.3 for the detail). For MARG, we only report scores for weaknesses because it only generates critiques of papers.

Metrics. We employed two types of metrics: focus similarity and text similarity, used in prior work (Zhou et al., 2024; Chamoun et al., 2024;

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

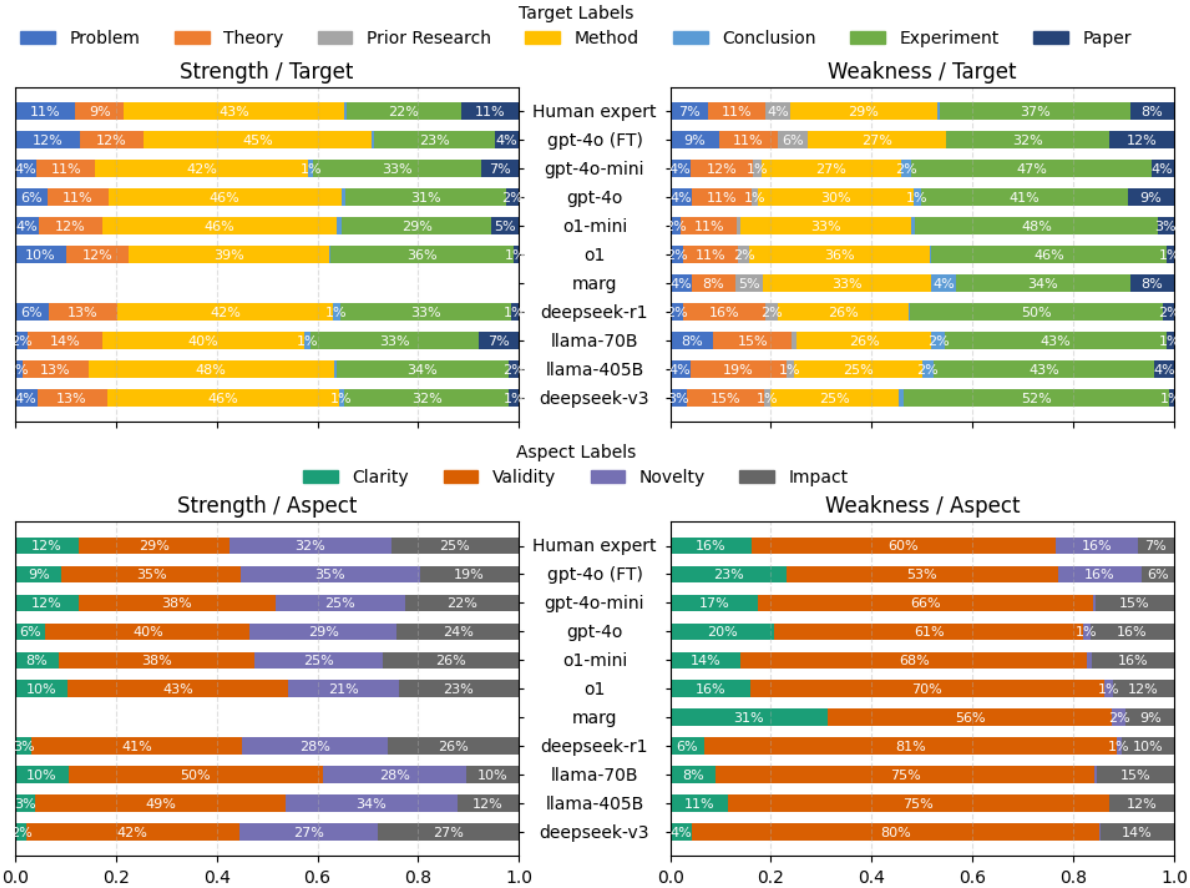

Figure 4: A visualization of focus distributions by target/aspect and strength/weakness, in a descending order of cosine similarity. Overall, both groups showed similar view points in reviewing papers, focusing on technical targets (i.e., Method, Experiment, and Theory) and validity. However, LLMs showed a more biased focus towards the technical validity whereas human experts exhibited more balanced focus. Moreover, all the LLMs lack consideration of Novelty for weaknesses compared to human experts, which is a significant limitation in reviewing papers.

Gao et al., 2025). For focus similarity, We measured Kullback-Leibler (KL) Divergence between the focus distributions of the models and human experts. We also measured F1 scores over the set of annotated (target, aspect) pairs as an agreement on review points. For text similarity, we measured ROUGE-L, BERTScore, and BLEU-4 between the LLM and expert reviews.

5.2 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). 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. The average review length of human experts and the models were 2639.76 and 3976.25, respectively. By comparing their focus distributions, we report the following key findings.

The fine-tuned `gpt-4o` produced focus distributions most closely aligned with that of human experts, while other models excelled in different evaluation dimensions. Table 3 shows the overall performance of the models. `gpt-4o (FT)` showed the highest focus distribution similarity, `DeepSeek-R1` achieved the best agreement on (target, aspect) labels, and `Llama-405B` showed the highest text similarity score. `gpt-4o` showed balanced performance, with moderate scores for both focus and text similarity. The results indicate the multifaceted nature of the paper review evaluation task. In other words, assessing the quality of reviews needs a holistic approach that integrates mul-

tiple and complementary metrics.

Overall, LLMs do not effectively identify key targets and aspects when reviewing papers. Table 3 shows the overall focus similarity and text similarity. The highest overall F1 score among the LLMs was 0.373, 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 is significantly different between human experts and LLMs. In general, LLMs showed higher recall (0.402) than precision (0.300) 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.

While overall agreement is low, both groups have similar primary focus in reviewing papers. Figure 4 shows a visualization of focus distributions between LLMs and human experts. 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 than Experiment in the strengths, but criticized Experiment more than Method in the weaknesses. For aspects, both groups considered Validity as the primary focus when identifying weaknesses. However, human experts focused more on Novelty in strengths whereas LLMs maintained Validity as the primary focus. For both groups, Impact received more attention in the strengths than weaknesses, whereas Clarity showed the opposite.

LLMs *consistently* exhibited a more biased focus, notably overlooking *novelty assessment* in identifying weaknesses. Although both groups had the similar primary focus, LLMs tend to concentrate on a few specific dimensions. For instance, for targets, LLMs focused primarily on Method and Experiment, with less focus on Prior Research (e.g., whether the paper adequately addresses prior work in positioning) and Problem (e.g., whether the task needs community attention) compared to human experts (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. The LLMs' biased focus was observed for *accepted* papers too, mostly criticizing experimental validity (See Appendix A.5). Notably, LLMs rarely focused on Novelty aspect in identifying weaknesses. This is a significant drawback, as a paper review requires a critical examination of novelty, by comparing them against existing work. Fortunately, we observed that `gpt-4o (FT)` identifies Novelty aspect in the weakness, as close as human experts.

Due to their biased focus, the level of agreement between LLMs and human experts varied across different labels. For targets and aspects that LLMs primarily focus on — Method (0.731, an average F1 score) and Experiment (0.671) targets and Validity (0.771) aspect — LLMs had a much higher level of agreement with human experts compared to other targets (0.213) and aspects (0.340). In the case of Experiment, the F1 score was consistently higher for weaknesses (0.835) than strengths (0.513), suggesting that LLMs are more effective at identifying concerns (e.g., lack of baselines or scope of evaluation) than strong points of experiments (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 (0.126). See Appendix A.4 for the full results.

LLMs showed similar patterns in their focus, 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).

6 Discussion

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

There is significant room for improving alignments between human experts and LLMs in paper reviewing. Our results show that LLMs exhibit a more biased focus, primarily assessing technical validity without contextual consideration,

compared to human experts. While fine-tuning yielded closer focus with human experts, the alignment of review points remained low. Since our focus-level evaluation only considered the target and aspect labels rather than their actual contents, we suspect that a more significant gap lies in the actual content addressed in the review items. For instance, even if two review points share the same label set (Experiment, Validity), they could point out different points such as lack of necessary baselines or lack of ablation studies to justify authors' arguments. Content-level investigations based on annotated facets may reveal more specific limitations of LLMs in reviewing papers, ultimately contributing to improving their reasoning capability.

Research should investigate the task of assessing the novelty of academic papers. Our finding illustrated that all untuned 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 their novelty assessment capability. We believe that creating the benchmark is a valuable contribution to the field, allowing LLMs to learn how to assess similarities between papers. Leveraging data in OpenReview could be an initial step as it contains experts' judgment on novelty of the paper for both positive and negative decisions.

A focus-level evaluation framework can offer unique value for guiding LLM training. The automated focus-level evaluation pipeline enables continuously tracking and evaluation of how LLMs focus on key facets of a paper over time, which aligns with the goals of holistic evaluation benchmarks (Liang et al., 2022; Srivastava et al., 2022). Beyond the language model evaluation, focus-level supervision can be incorporated during the training process; reward functions can be designed to encourage balanced focus aligned with human experts or even purposefully facilitate a certain focus (e.g., building a novelty-focused reviewer) (Yang et al., 2024; Agnihotri et al., 2025). Furthermore, the framework is generalizable to other domains where the output spans multiple facets—such as debating, decision making, and educational feedback—making *focus* a critical factor in generated outputs.

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., 2024; 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 a framework for focus-level evaluation of LLM reviews, which systematically analyzes where LLM reviews direct their praise and criticism based on pre-defined facets. Our findings suggest that LLMs need to adopt a more balanced perspective, have higher agreement with human experts about the target and aspect in the strengths and weaknesses, and place greater emphasis on novelty assessment when criticizing papers. We believe that the focus-level evaluation can contribute to ongoing evaluation of LLMs' paper review capabilities within the rapid pace of LLM developments.

Limitation

This paper has the following limitations. First, our dataset focuses solely 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 (0.80) 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 facets only considers two dimensions, which may undervalue the unique contributions of LLM-generated reviews.

References

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. 2025. Multi-objective preference optimization: Improving human alignment of generative models. *Preprint*, arXiv:2505.10892.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

Xian Gao, Jiacheng Ruan, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Reviewagents: Bridging the gap between human and ai-generated paper reviews. *ArXiv*, abs/2503.08506.

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. 2024. 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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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.

Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying aspects in peer reviews. *arXiv preprint arXiv:2504.06910*.

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.

Lakshmi Ramachandran, Edward F Gehringer, and Ravi K Yadav. 2017. Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, 27(3):534–581.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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 Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei 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 4. An item refers to a specific requirement mentioned in the guidelines, which serves as a distinct criterion for reviewing or writing a paper.

Table 4: 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**

[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

A.2.2 Target and aspect facets

Table 5: We aim to analyze focus distributions of LLM reviews based on the targets and aspects. To identify the specific facets 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). The facets were used as the codebook for human annotations.

Target	
Facet	**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	
Facet	**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.

A.2.3 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.

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.2.4 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.

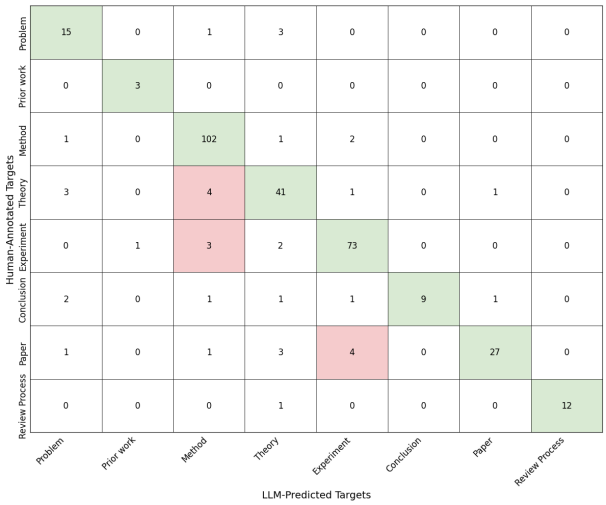

Figure 13: LLM vs. human target annotation

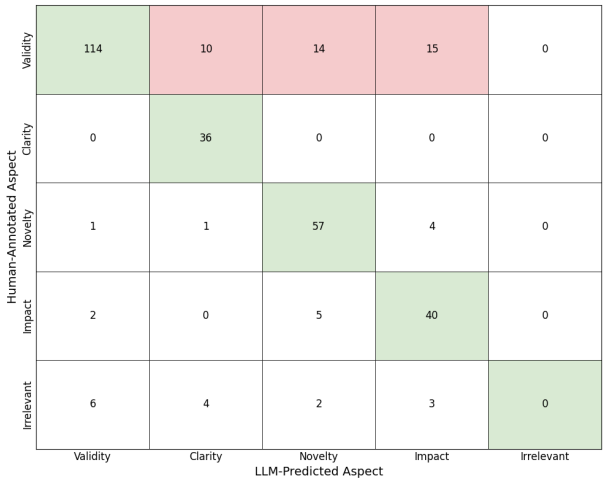

Figure 14: LLM vs. human aspect annotation

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.3 Fine-Tuning Details

A.3.1 Fine-Tuning Dataset Construction

We constructed the fine-tuning dataset based on the corpus of papers described in Section 3. We retained 582 training samples and 98 test samples. 5 samples were excluded during tokenization due to exceeding the model's maximum token length

A.3.2 Fine-Tuning Method

We employed supervised fine-tuning (SFT) to adapt the GPT-4o base model to our task-specific objectives. Fine-tuning was conducted using the OpenAI Fine-Tuning API[6], which abstracts away hardware and infrastructure details. Therefore, we do not report GPU type or compute hours. Table 6 summarizes the hyperparameter configuration used during training.

Table 6: Hyperparameter settings for supervised fine-tuning.

Parameter	**Value**
total epochs	4
batch size	4
learning rate multiplier	0.1

A.4 Detailed Evaluation 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

[6] https://platform.openai.com/docs/api-reference/fine-tuning

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

A.5 Results using accepted papers

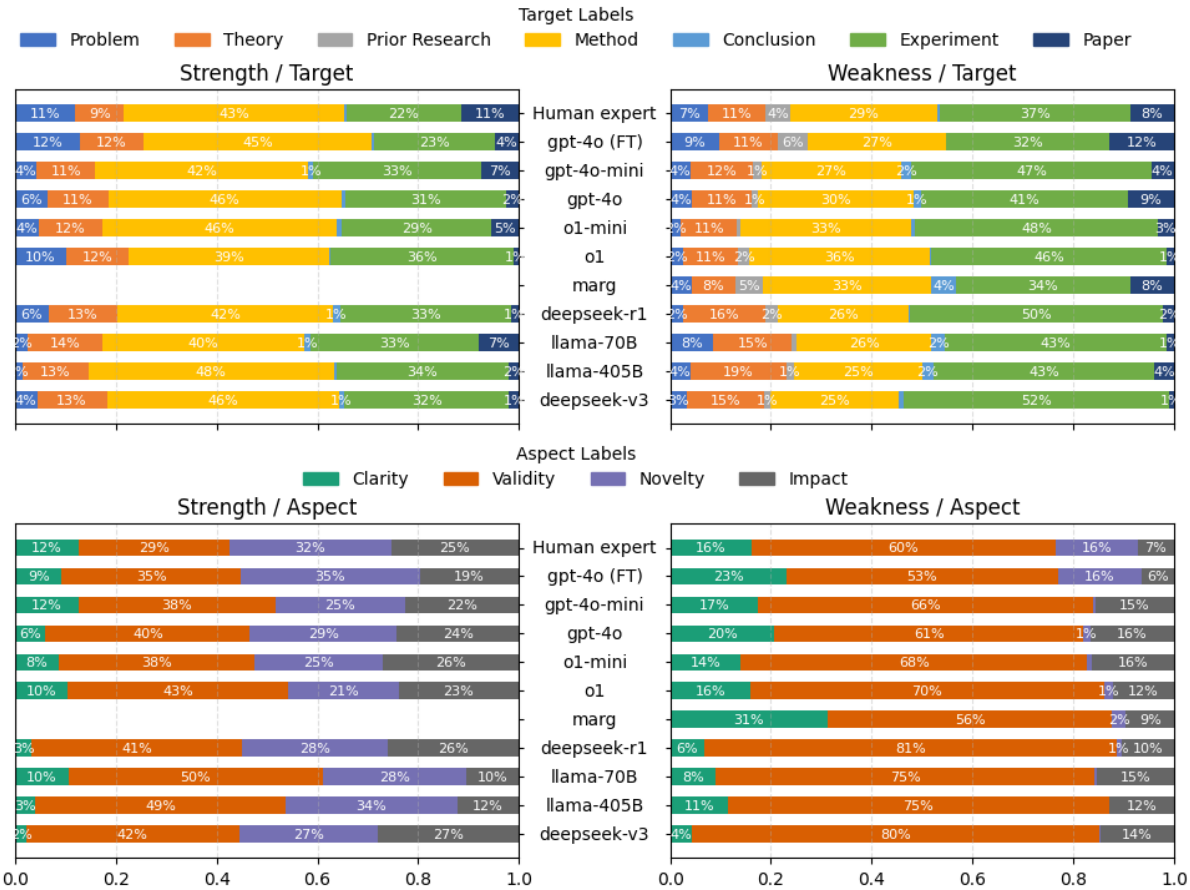

Figure 17: A visualization of focus distributions by target/aspect and strength/weakness for LLMs and human experts using *accepted* papers, in a descending order of KL divergence. We observed a few notable differences in the pattern, compared to the evaluation results using rejected papers. First, there exists a much larger gap in the Weakness-Experiment, meaning that human experts criticize experiments significantly less than LLMs. In strengths, human experts mostly praise Novelty and Impact than Validity, but LLMs tend to praise the Validity the most. We observed the same pattern in Weakness-Novelty, meaning that LLMs neglect the novelty aspect in criticizing the papers.