WarrantScore: Modeling Warrants between Claims and Evidence for Substantiation Evaluation in Peer Reviews

Anonymous ACL submission

Abstract

The scientific peer-review process is facing a shortage of human resources due to the rapid growth in the number of submitted papers. The use of language models to reduce the human cost of peer review has been actively explored as a potential solution to this challenge. A method has been proposed to evaluate the level of substantiation in scientific reviews in a manner that is interpretable by humans. This method extracts the core components of an argument, claims and evidence, and assesses the level of substantiation based on the proportion of claims supported by evidence. The level of substantiation refers to the extent to which claims are based on objective facts. However, when assessing the level of substantiation, simply detecting the presence or absence of supporting evidence for a claim is insufficient; it is also necessary to accurately assess the logical inference between a claim and its evidence. We propose a new evaluation metric for scientific review comments that assesses the logical inference between claims and evidence. Experimental results show that the proposed method achieves a higher correlation with human scores than conventional methods, indicating its potential to better support the efficiency of the peer-review process.

1 Introduction

The rapid development of scientific research has raised concerns about the increasing human costs associated with the peer-review process. Leading international conferences in computer science have seen a sharp increase in the number of papers submitted year-by-year. However, the number of reviewers has not matched the increase in the number of papers submitted (Tran et al., 2020). Due to the shortage of reviewers, the burden on conference organizers of ensuring the quality of reviews continues to increase (Stelmakh et al., 2021).

A promising approach to address these issues is language model-assisted review. For example, automatic detection of low-quality reviews by language models may encourage reviewers to improve their reviews or support meta-reviewers in making decisions. For these purposes, it is important not only to assign a quality score to a review, but also to provide an evaluation process in a form that is easily understandable to humans.

As an evaluation metric that is easily understood by humans, SubstanScore (Guo et al., 2023) has been proposed. SubstanScore evaluates the level of substantiation by extracting claims and evidence from reviews and measuring the proportion of claims supported by evidence. SubstanScore not only provides the final evaluation score but also generates descriptions identifying the claims and supporting evidence within the review. Descriptions of claims and evidence allow for the assessment of the level of substantiation in scientific reviews in a way that is interpretable by humans.

SubstanScore evaluates the level of substantiation by focusing only on claims and evidence, which are two of the three main components of Toulmin's model (Toulmin, 2003), a well-established argument structure. However, the presence of claim and evidence pairs alone does not ensure a comprehensive evaluation of the quality of review comments. For example, consider the claim "I will reject the submitted paper" paired with the evidence "The main topic of the paper has already been thoroughly examined in many studies." The logical inference between claims and evidence within a review is not always sufficient because the remaining core element of Toulmin's model, the warrant such as "topics on which many studies have been conducted are not accepted" which serves to logically connect the claim and the evidence, may be somewhat credible but is not always entirely convincing. Therefore, evaluating the level of substantiation requires not only checking for the presence of supporting evidence

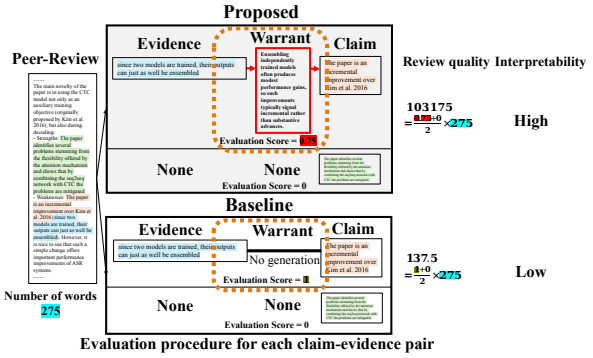

Figure 1: Comparison between proposed peer-review evaluation metrics that evaluate claims, evidence, and warrants and baseline metrics that only evaluate claims and evidence.

but also accurately assessing the causal connection between the claim and its supporting evidence.

Inspired by Toulmin's model, we proposed a new evaluation metric for scientific review comments called WarrantScore as an extension of SubstanScore alone, by explicitly generating and evaluating warrants. Figure 1 illustrates this concept.

We compared WarrantScore and SubstanScore using datasets annotated with subjective evaluations of humans of the level of substantiation, namely SubstanReview (Guo et al., 2023) and RottenReview (Ebrahimi et al., 2025). The comparison was based on the correlation with subjective evaluations of humans. The results demonstrated that WarrantScore exhibits a stronger correlation with subjective evaluations of humans than SubstanScore. Moreover, we confirmed the superiority of WarrantScore over regression models using subjective evaluations of humans as the target variable. To support the superiority of WarrantScore, we analyzed the ability of the evaluation metrics to assess the quality of scientific reviews, regardless of superficial linguistic features such as the number of words in a review. Each experiment is described in Section 5 and Section 6.

Our research addresses the challenge of automatically assessing the quality of scientific reviews in a way that is independent of superficial linguistic features and interpretable to humans.

2 Related Work

In this section, we first discuss the existing automatic evaluation metrics for scientific peer review comments in Section 2.1 and describe the automatic extraction of argument components in Section 2.2.

2.1 Evaluation of Peer Review Comments

Various evaluation metrics targeting specific aspects of scientific review comments have been proposed. For example, some focus on aspects such as harshness (Verma et al., 2022), thoroughness and helpfulness (Severin et al., 2022), or comprehensiveness (Yuan et al., 2022). However, these metrics are regression-based approaches that aim to predict subjective evaluations of human as the target variable. Therefore, even if these metrics provide an evaluation score for a given scientific review comment, humans cannot understand why that specific score was assigned.

SubstanScore attempts to automatically extract claims and evidence from scientific review comments and evaluate the level of substantiation based on the proportion of claims supported by evidence. Unlike regression methods, SubstanScore not only provides the final evaluation score but also shows, in a way that humans can verify, how the score was computed through the evaluation process. Whether the extracted evidence is truly correct and whether the evidence logically supports the claims remains unverified. Ryu et al. (Ryu et al., 2025) proposed extracting the argumentative structure of scientific reviews using large language models (LLMs) to detect erroneous criticisms and verified the factual correctness of the evidence using LLMs. However, in the context of scientific review evaluation, no established method exists even now to verify that the evidence logically supports the claims.

2.2 Argument Mining

An attempt has been made in the general domain to automatically extract argumentative components, such as warrants, from text using LLMs (Gupta et al., 2024). In their study, GPT-4 (Achiam et al., 2023) was prompted to generate argument components of Toulmin's model by adding the phrase "According to Toulmin model" to the input text. As a result, the warrants generated by GPT-4 were rated acceptable by two domain experts in 61.7% of cases. They consider a warrant acceptable if it is: a) relevant and fully explains the link between the claim-reason pair, b) not trivial, c) must hold for the claim to be inferred from the reason, even if it does not align with the annotator's or reader's personal beliefs.

Since two experts rated 45.5% of the warrants generated by humans as acceptable, it is

suggested that sufficiently parameterized LLMs such as GPT-4 can generate warrants that humans may evaluate more favorably than human-authored warrants. Their study also has made available a dataset includes a total of 450 warrants along with binary acceptability labels assigned by two experts. The insights and resources released demonstrate the potential to generate and evaluate implicit warrants in scientific reviews.

3 Problem Formulation

In this section, we formalize the task of generating warrants, defined as text that establishes the logical inference between claims and evidence, which our proposed method is designed to address.

First, the set of claims extracted from a scientific review comment by humans or LLMs is defined as $C = \{c_1, c_2, \ldots, c_N\}$, and similarly, the set of supporting evidence for these claims is defined as $E = \{e_1, e_2, \ldots, e_N\}$. Each c and e represents text indicating a claim and its corresponding evidence within the scientific review comment. Elements of the set C are all non $\emptyset$, but elements of E may be $\emptyset$, reflecting that in scientific review comments, a claim does not always have supporting evidence.

Let w be defined as text that logically connects c and e; this text is called the warrant. The set of warrants is $W = \{w_1, w_2, \ldots, w_N\}$. When e is $\emptyset$, then w is also $\emptyset$. Let LM denote the LLM as a function and let p denote the instruction prompt. The relationship among C, E, and W is formalized in Equation (1).

$$\begin{aligned} W &= \{w_1, w_2, \cdots, w_N\} \\ &= \big\{\mathrm{LM}(p, c_1, e_1),\ \mathrm{LM}(p, c_2, e_2), \\ &\qquad \ldots,\ \mathrm{LM}(p, c_N, e_N)\big\}. \end{aligned} \tag{1}$$

Our proposed method assesses the level of substantiation in a scientific review comment by evaluating W generated by LLM.

4 Proposed Method

In this section, we first describe the baseline method in Section 4.1, then define the proposed method in Section 4.2 and finally detail the algorithm of our proposed method in Section 4.3.

4.1 SubstanScore (Baseline)

Guo et al. (Guo et al., 2023) proposed SubstanScore, an evaluation metric for the level of substantiation in scientific review comments. SubstanScore is calculated by multiplying the proportion of claims with supporting evidence by the number of words in the review. The definition of SubstanScore is shown in Equation (2).

$$\begin{aligned} &\text{SubstanScore} \\ &= \frac{\sum_{n=1}^{N} \text{support_score}(c_n, e_n)}{|C|} \\ &\times \text{len(review)} \\ &= \text{supported_claims} \times \text{len(review)}. \end{aligned} \tag{2}$$

len(review) indicates the number of words in a scientific review comment, and support_score(c, e) is a function that returns 0 if e is $\emptyset$, and 1 otherwise. supported_claims represents the proportion of claims supported by evidence, calculated as

$$\begin{aligned} &\text{supported_claims} \\ &= \frac{\sum_{n=1}^{N} \text{support_score}(c_n, e_n)}{|C|}. \end{aligned} \tag{3}$$

SubstanScore does not verify whether the supporting evidence truly justifies the claim.

4.2 WarrantScore

We proposed WarrantScore, an evaluation metric for scientific review comments that generates warrants, which are texts that logically connect claims and supporting evidence, and assess the plausibility of warrants, as an extension of SubstanScore. The definition of WarrantScore is expressed in Equation (4).

$$\begin{aligned} &\text{WarrantScore} \\ &= \frac{\sum_{n=1}^{N} \text{support_score}(c_n, e_n, w_n)}{|C|} \\ &\times \text{len(review)} \\ &= \frac{\sum_{n=1}^{N} \mathrm{V}(w_n) \times \text{support_score}(c_n, e_n)}{|C|} \\ &\times \text{len(review)} \\ &= \text{warrant_rate} \times \text{len(review)}. \\ &\quad * \text{ note that } w_n = \mathrm{LM}(p, c_n, e_n). \end{aligned} \tag{4}$$

WarrantScore evaluates both the presence of supporting evidence for claims and the plausibility of the warrant connecting each claim–evidence pair. In WarrantScore, support_score(c, e, w) is a

function that returns 0 if $e = \emptyset$, and otherwise returns an evaluation score for w in the range from 0 to 1. The warrant_rate represents the mean evaluation score of warrants, calculated as

$$\text{warrant_rate} = \frac{\sum_{n=1}^{N} \mathrm{V}(w_n) \times \text{support_score}(c_n, e_n)}{|C|}. \quad (5)$$

SubstanScore can be interpreted as a special case of WarrantScore, in which $\mathrm{V}(w)$ is always 1 when $\text{support_score}(c, e) = 1$, or, alternatively, by replacing $\text{support_score}(c_n, e_n) \in \{0, 1\}$ with $\text{support_score}(c_n, e_n, w_n) \in [0, 1]$.

4.3 Warrant Generation and Evaluation

The pipeline for generating warrants from claims and evidence, and evaluating these warrants for the calculation of WarrantScore, is illustrated in Figure 2. In our study, a warrant is defined as text that expresses implicit common sense and satisfies the acceptability criteria to connect a claim and its evidence. Since the definition of warrants in Toulmin's model is highly complex, our study focused on generating warrants that are central to evaluation. Specifically, we only focused on the common sense that reviewers have accumulated through their previous reviews or research experience. Details of the implementation for warrant generation and evaluation can be found in Appendix A and Appendix B, respectively.

We use LLM to generate warrants that connect claims and evidence. Then, another LLM classifies the generated warrant as either acceptable or not acceptable using binary classification; if not acceptable, this process is repeated up to three times. If the warrant remains unacceptable after the third classification, we set $w = \emptyset$.

In our study, warrant evaluation is defined as assessing the plausibility of w as $\mathrm{V}(w)$. Using an LLM-as-a-judge, the plausibility of each w is evaluated on a four-point Likert scale. The warrant score $\mathrm{V}(w)$ is derived from a four-point Likert rating $v \in \{1, 2, 3, 4\}$, with the value linearly scaled to the range $[0, 1]$ by multiplying by 0.25. In the same way as SubstanScore, if e is $\emptyset$, WarrantScore assigns $\mathrm{V}(w) = 0$. If $e \neq \emptyset$ but $w = \emptyset$, then $V(w) = 1$, following the SubstanScore assumption that any claim supported by evidence is logically connected.

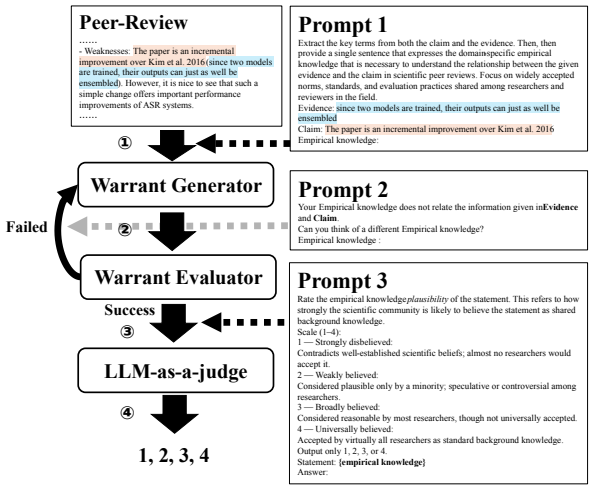

Figure 2: Warrant-generation and evaluation pipeline using LLMs: first, an LLM (warrant generator) generates a warrant that connects a claim and evidence pair; next, a binary classification model (warrant evaluator) verifies its acceptability; and finally, another LLM (LLM-as-a-judge) evaluates the warrant on a four-point scale.

5 Correlation Analysis with Human Judgments

In this experiment, we verified that WarrantScore accurately generates warrants that connect claims and evidence, and that it exhibits a stronger correlation with the level of substantiation annotated by humans than SubstanScore.

5.1 Evaluation Metrics

We categorize the evaluation metrics for scientific review comments into our proposed methods and baselines. Our proposed methods, WarrantScore and warrant_rate, are described in Equation (4) and Equation (5), respectively. For comparison, we also consider the following baseline methods: SubstanScore and supported_claims, which are defined in Equation (2) and Equation (3), respectively. We further consider:

- coherence_rate: The mean semantic similarity, defined as $\frac{\sum_{n=1}^{N} \mathrm{Sem}(c_n, e_n) \times \text{support_score}(c_n, e_n)}{|C|}$, used to assess the coherence between c_n and e_n. (Lapata and Barzilay, 2005). $\mathrm{Sem}(c, e)$ indicates the calculation of the semantic similarity between c_n and e_n by Sentence-BERT (Reimers and Gurevych, 2019) *.
- random50%_rate: The 50th percentile of the random evaluation that correlate human eval-

*sentence-transformers/all-mpnet-base-v2

uastion score, calculated by performing the operation $\frac{\sum_{n=1}^{N} r_n \times 0.25 \times \text{support_score}(c_n, e_n)}{|C|}$ twenty thousand times for all warrants in a dataset, where r is an independently sampled uniform random variable from {1, 2, 3, 4} for each n.

- CoherenceScore: A simple multiple coherence rate and number of words in a review, defined as coherence_rate × len(review).
- Random50%Score: Analogous to random50%_rate, but calculated using the formula $\frac{\sum_{n=1}^{N} r_n \times 0.25 \times \text{support_score}(c_n, e_n)}{|C|} \times \text{len}(\text{review})$, where $\text{len}(\text{review})$.
- RM: A regression model based on Llama has been trained using subjective evaluations of the level of substantiation annotated by human. Details of the implementation can be found in Appendix C.
- review_len: The number of words in a review, defined as len(review).
- flesch reading ease (fre) (Flesch, 1948): a measure of readability given by $\text{FRE} = 206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}}\right)$.
- type token ration (ttr) (Richards, 1987): measure of vocabulary diversity defined as $\text{TTR} = \frac{V}{\text{total words}}$, where V is the size of vocabulary within the review.

The coherence_rate and random50%_rate were used to demonstrate that the correlation between WarrantScore and subjective evaluations of humans is not attributable to chance. The RM was used as an example of a regression model trained with subjective evaluations of humans as the target variable, to compare with the scientific review evaluation metrics based on the argumentation structure. To show that changing the warrant evaluation model does not substantially affect the performance of warrant_rate and WarrantScore, we used two LLM, namely Gemini 2.5 Flash and GPT-5 for warrant evaluation. We also present simple linguistic features, such as review_len, fre, and ttr, that correlate with aspects of human subjective evaluations of scientific review comments. These features are used to highlight the differences between our proposed scientific review evaluation metric and general linguistic properties.

5.2 Dataset

Toward reliable correlation analysis, we used two datasets that were annotated by humans for the level of substantiation: SubstanReview and RottenReview. An overview of each dataset is provided below.

SubstanReview (Guo et al., 2023) is a dataset in which 550 scientific review comments from the natural language processing scientific review dataset NLPeer (Dycke et al., 2023) are annotated by humans for claims and evidence. Of these, 50 reviews have been assigned the level of substantiation on a three-point Likert scale by three human annotators. In this experiment, only the 50 reviews annotated by humans for their level of substantiation were used to evaluate the proposed evaluation metrics. We aggregated the levels of substantiation assigned by the three annotators by taking the maximum value to reduce variability in scores among human annotators, as evaluating scientific review quality is a challenging task. RM for SubstanReview was not calculated, since the dataset size was insufficient for reliable model training and evaluation, making correlation estimates unstable.

RottenReview (Ebrahimi et al., 2025) contains 753 scientific review comments randomly sampled from *NeurIPS*, *ICLR*, *F1000*, and *SWJ*. Multiple human annotators evaluated these reviews on a five-point Likert scale across several aspects, including Objectivity, which indicates "the presence of unbiased, evidence-based commentary." Objectivity is a similar aspect to substantiation in SubstanReview. The annotation scores were consolidated by taking the maximum score, because the different annotators evaluated the same review. After deduplication, the number of unique reviews annotated by humans totaled 509 reviews. For this experiment, the correlation between Objectivity and various review evaluation metrics was used to evaluate the metrics. To extract the claim and evidence for all reviews in RottenReview, the state-of-the-art BERT-based model trained on SubstanReview, ModernBERT (Warner et al., 2025), was used. Details of the claim and evidence tagging implementation are provided in Appendix D.

5.3 Result

We evaluated the evaluation metrics for the scientific review by computing Spearman's rank correlation coefficient between subjective evaluations

Table 1: An example of generated warrant in SubstanReview.

Claim	It's great that they also report results on another language,
Evidence	showing large improvements over existing work on Japanese CCG parsing.
Warrant	When a parsing method produces large gains on Japanese CCG and comparable gains on another language, it usually signals language-agnostic modeling improvements rather than dataset-specific tricks.

Table 2: Human correlation with peer-review evaluation feature in SubstanReview and RottenReview. The warrants of warrant_rate and WarrantScore is evaluated by GPT-5.

Metric	SubstanReview	RottenReview
review_len	**0.77**	**0.33**
fre	-0.09	0.19
ttr	-0.73	-0.31
supported_claims	0.45	0.10
coherence_rate	0.67	**0.18**
warrant_rate	**0.69**	**0.18**
SubstanScore	0.70	0.25
CoherenceScore	0.79	0.25
WarrantScore	**0.82**	**0.27**
RM	N/A	**0.49**

of human and the evaluation metric scores. Table 2 shows the correlation coefficients between them in SubstanReview and RottenReview. In SubstanReview, both warrant_rate and WarrantScore showed higher correlations with subjective evaluations of humans compared to supported_claims and SubstanScore, respectively. This suggests that evaluating warrants aligns more closely with human judgments. warrant_rate and WarrantScore are more strongly correlated with human evaluations than coherence_rate and CoherenceScore, respectively, suggesting that the explicit generation of warrants leads to better alignment with human judgments. The highest correlation coefficient in SubstanReview was observed for WarrantScore, indicating that when claims and evidence are annotated by humans, WarrantScore most accurately assesses the level of substantiation in scientific reviews.

Similarly, in RottenReview, warrant_rate and WarrantScore as well as coherence_rate and CoherenceScore, showed high correlations with subjective evaluations of humans, outperforming supported_claims and SubstanScore, respectively. Unlike the results with SubstanReview, warrant_rate and WarrantScore in RottenReview displayed correlation coefficients with subjective evaluations of humans nearly equivalent to correlation coefficients of coherence_rate and CoherenceScore, which are weaker than the correlation between review_len and subjective evaluations of humans. Although WarrantScore is more strongly correlated with human subjective evaluations than SubstanScore, the correlation coefficients of both metrics with human evaluations remain relatively low.

Table 3: Human correlation with warrant_rate and WarrantScore in SubstanReview and RottenReview. The intersection of the GPT-5 and the rate represents the warrant_rate evaluated by GPT-5.

Model	SubstanReview		RottenReview	
	rate	score	rate	score
random50%	0.68	0.79	0.09	0.19
Gemini 2.5 Flash	**0.69**	0.80	0.16	0.26
GPT-5	**0.69**	**0.82**	**0.18**	**0.27**

In RottenReview, the evaluation method that showed the highest correlation with subjective evaluation of humans was RM, a regression model. In RottenReview, the assigned scores are not explicitly linked to claim-evidence pairs, and the claim-evidence pairs used in our experiment were automatically extracted. This result highlights a limitation of our approach, which does not directly regress on review scores. However, since our method does not require labeled training data and the evaluation dataset for RM covers only 30% of the RottenReview, we cannot conclude that our method is inherently inferior.

Table 3 shows the correlation coefficients of warrant_rate, WarrantScore, and random50% for each LLM. In both SubstanReview and RottenReview, when evaluating warrants with LLMs, these metrics exceeded the correlation coefficients obtained from random50%. The increased correlation with human subjective assessments, observed when LLMs evaluate warrants more accurately than random, suggests that proper warrant evaluation is crucial for our proposed method.

An example of claims, evidence, and warrants generated by LLMs in SubstanReview is shown in Table 1. Intuitively, in SubstanReview, where claims and evidence are annotated by humans, LLMs appear to generate warrants that connect the given claims and evidence as common sense, functioning as plausible justifications for claims.

6 Robustness Analysis of Evaluation Metrics

In this section, we analyzed the advantages of the proposed method, particularly its robustness, by comparing it with a regression model, since the correlation between our proposed method and subjective evaluations of humans is still relatively low. We confirmed that each evaluation metric tends to assign higher scores to good reviews than to bad ones, regardless of the number of words in a review. This property is a desirable characteristic of evaluation metrics for peer-review.

Goldberg et al. (Goldberg et al., 2025) analyzed the reliability of scientific review comment evaluations by presenting experts with review comments submitted to international conferences, along with artificially elongated versions of those reviews that retain the original content but include redundant text. The results indicated that experts tend to evaluate elongated reviews more favorably than the original reviews. This highlights the need for evaluation metrics that accurately evaluate the quality of scientific reviews while being unaffected by biases introduced by number of words in a review, unlike subjective evaluations of human.

The extent to which review evaluation metrics are influenced by number of words in a review was assessed using the following two datasets.

One dataset confirms that scientific review evaluation metrics tend to assign lower scores to deficient reviews compared to sufficient reviews. This dataset is referred to as ReviewGuardReview. ReviewGuard (Zhang et al., 2025) is a framework that uses an LLM-as-a-judge to evaluate the sufficiency of reviews, labeling reviews as sufficient or deficient. ReviewGuard defines a review as sufficient if it meets all three of the following criteria: a) commitment (the extent of substantive engagement with the manuscript), b) constructive intention (a fair-minded desire to improve the work), c) domain knowledgeability (sufficient expertise to make an informed judgment). Missing any of them makes it deficient.

In their study, following the ReviewGuard framework, labels were assigned to all reviews of *ICLR 2019* using GPT-OSS. Subsequently, 500 reviews labeled as sufficient and 500 reviews labeled as deficient were randomly sampled to construct the dataset. Note that only the review texts were used as input; the paper's title and abstract, which are included in the original framework, were omitted, as the review evaluation metrics do not reference these details. In our study, we assume that sufficient reviews are well substantiated, while deficient reviews are poorly substantiated.

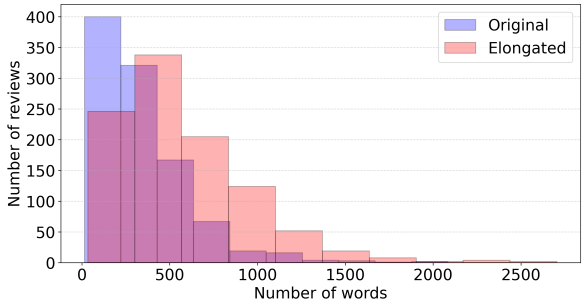

Figure 3: Number of words in the original peer-review and the elongated peer-review.

Table 4: The mean evaluation score of each feature (e.g., warrant_rate) for peer-review comments labeled as sufficient or deficient reviews in ReviewGuardReview. The warrants of warrant_rate and WarrantScore is evaluated by GPT-5.

Metric	sufficient	deficient	d(↑)	p(↓)
review_len	**366.78**	308.16	**0.22**	**0.00**
fre	**41.72**	41.25	0.04	0.26
ttr	0.55	**0.59**	-0.37	1.00
supported_claims	0.44	**0.46**	-0.05	0.81
coherence_rate	**0.11**	0.09	**0.21**	**0.00**
warrant_rate	**0.28**	0.23	0.18	**0.00**
SubstanScore	**175.09**	150.07	0.14	**0.01**
CoherenceScore	**48.11**	38.38	**0.17**	**0.00**
WarrantScore	**115.38**	95.25	0.16	**0.01**
RM	1.59	**1.77**	-0.15	0.99

Another dataset is used to verify whether scientific review evaluation metrics assign similar scores to original and elongated reviews. This dataset is referred to as ElongatedReview. ElongatedReview is inspired by the experiment conducted by Goldberg et al. (Goldberg et al., 2025), which showed that experts tend to rate elongated reviews more highly. ElongatedReview is created by taking 1,000 reviews from ReviewGuardReview and generating elided summaries of each review using GPT-OSS, then appending these summaries to the original reviews. Consequently, for each of the 1,000 reviews, a nearly identical but more elongated version was produced. Figure 3 shows the word counts for both the original and elongated reviews. The average word count of the original reviews is 337.47 words, with the minimum being 15 words and the maximum 2,084 words. For the elongated reviews, the average word count is 572.26 words, with the minimum at 34 words and the maximum at 2,705 words.

Table 5: The mean evaluation score of each feature (e.g., warrant_rate) for peer-review comments labeled as original and elongated peer reviews in ElongatedReview. The warrants of warrant_rate and WarrantScore is evaluated by GPT-5.

Metric	original	elongated	d (↓)	p(↑)
review_len	337.47	**572.26**	1.99	0.00
fre	**41.49**	31.96	**-1.84**	0.00
ttr	**0.57**	0.47	-2.17	0.00
supported_claims	0.45	**0.47**	**0.07**	**0.03**
coherence_rate	0.10	**0.11**	0.11	0.00
warrant_rate	0.26	**0.27**	**0.07**	**0.02**
SubstanScore	162.58	**284.66**	0.89	0.00
CoherenceScore	42.34	**72.12**	0.69	0.00
WarrantScore	105.32	**174.91**	**0.68**	0.00
RM	**1.68**	1.05	-0.77	0.00

6.1 Evaluation Metric

In the evaluation using ReviewGuardReview, a one-sided independent t-test was conducted with the null hypothesis that the sufficient reviews are not rated higher than the deficient reviews. If the null hypothesis is rejected, it indicates that the sufficient reviews are rated higher than the deficient ones. In ElongatedReview, a paired t-test was used with the null hypothesis that the evaluation scores of the original reviews and the elongated reviews are the same. If the null hypothesis is not rejected, it suggests that the original and elongated reviews are evaluated as equivalent. Both tests in the dataset are significant at the 1% level. The effect sizes for each comparison were calculated using Cohen's d (Cohen, 2013). For ReviewGuardReview, d was computed as the difference between sufficient and deficient reviews (sufficient − deficient). For ElongatedReview, d was computed as the difference between the evaluation scores of the original reviews and the elongated reviews (original − elongated).

For the calculation of WarrantScore and SubstanScore, all review comments in datasets were automatically tagged with claims and evidence using ModernBERT. In this experiment, desirable evaluation metrics were those that assign higher scores to sufficient reviews compared to deficient ones, and these scores should be invariant with respect to the number of words in a review.

6.2 Result

The evaluation results of the scientific review metrics in ReviewGuard are shown in Table 4. The null hypothesis that sufficient reviews are rated lower than deficient reviews was rejected for the following metrics: review_len, coherence_rate, warrant_rate, SubstanScore, CoherenceScore, and WarrantScore. These indicators tend to assign higher scores to sufficient reviews compared to deficient ones. In ReviewGuardReview, RM, the regression method based on the highest correlation coefficient with subjective evaluations of humans does not show a tendency to rate sufficient reviews higher than deficient reviews. Conversely, warrant_rate, which has limited correlation with subjective evaluations of humans, can still rate sufficient reviews higher than deficient ones. This suggests that the overall rankings of scientific review evaluation metrics cannot be determined solely based on their correlation with subjective evaluations of humans.

The evaluation results for scientific review evaluation metrics based on ElongatedReview are shown in Table 5. The null hypothesis that original reviews and elongated reviews are evaluated equally was not rejected for the following metrics: supported_claims and warrante_rate. These indicators tend to assign similar scores to original reviews and elongated ones. In other words, in terms of robustness, supported_claims and warrant_rate have an advantage over other features.

According to the experimental results, only the warrant_rate, a core component of WarrantScore, provides a robust evaluation of peer-review comments, as the null hypothesis that sufficient reviews are rated lower than deficient reviews was rejected, whereas the null hypothesis that original and elongated reviews are evaluated equally was not rejected.

7 Conclusion

In this study, we proposed a new framework to support the evaluation of scientific reviews, with a particular focus on claim–evidence pairs and their logical inferences. Specifically, we introduced WarrantScore, a metric for evaluating peer-review comments that considers not only supporting evidence but also the plausibility of warrants connecting claims and evidence. Experiments on human-annotated datasets show that WarrantScore correlates more strongly with human judgments than conventional metrics, indicating that modeling warrants enhances both alignment with human evaluations and the interpretability of review assessments.

Limitations

One limitation of our study is that we fixed the warrant, the rationale connecting claims and evidence, to a single form and limited the types of warrants to common sense likely acquired by reviewers through past reviews and research experience. Warrants are not necessarily unique, and their types are diverse. Consequently, the evaluation value of an arbitrary warrant is likely non-unique. In our proposed approach, we evaluated candidate warrants for each claim and evidence pair as a four-point Likert scale, using an LLM-as-a-judge. Although the generated warrants are understandable to humans, interpreting the evaluation scores assigned to these warrants remains challenging. Our primary demonstration is that, by leveraging the concept of warrants, generating and assessing implicit text that connect claims and evidence, we can improve the correlation between scientific review evaluation metrics and subjective evaluations of human, as well as the interpretability of the evaluation scores themselves.

Another limitation of our study concerns the small dataset size and the differences in domains. We used two publicly available human-annotated datasets, as well as two datasets generated with large language models. Our proposed method shows consistent performance improvements over conventional methods on all datasets, but the actual effectiveness remains unclear. Clarifying the practical performance of our approach requires further development of a large, human-annotated datasets of scientific reviews.

Ethics Statements

AI Assistants in our study including ChatGPT and Google Gemini, were used in accordance with the ACL Policy on AI Writing Assistance. We primarily used them to assist with coding and writing, but all code and text outputs were manually reviewed. The authors take full responsibility for all of them.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. Nlpeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 5049–5073.

Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2025. Rottenreviews: Benchmarking review quality with human and llm-based judgments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5642–5649.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. 2025. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PloS one*, 20(4):e0320444.

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

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 3982. Association for Computational Linguistics.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Hyun Ryu, Doohyuk Jang, Hyemin S Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, and 1 others. 2025. Reviewscore: Misinformed peer review detection with large language models. *arXiv preprint arXiv:2509.21679*.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3074–3080.

Anna Severin, Michaela Strinzel, Matthias Egger, Tiago Barros, Alexander Sokolov, Julia Vilstrup Mouatt, and Stefan Müller. 2022. Journal impact factor and peer review thoroughness and helpfulness: A supervised machine learning study. *arXiv preprint arXiv:2207.09821*.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4785–4793.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. 2020. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

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

Haoxuan Zhang, Ruochi Li, Sarthak Shrestha, Shree Harshini Mamidala, Revanth Putta, Arka Krishan Aggarwal, Ting Xiao, Junhua Ding, and Haihua Chen. 2025. Reviewguard: Enhancing deficient peer review detection via llm-driven data augmentation. *arXiv preprint arXiv:2510.16549*.

Appendix

A Detail of Warrant Generation and Evaluation

For the generation of warrants, we utilized GPT-OSS [†], a large open-source LLM, with the temperature set to 0. When regenerating warrants, the previously generated warrants are provided as context. During warrant evaluation, the maximum number of output tokens generated by the LLM was set to 2, and the parameter do_sample was set to false.

B Warrant Acceptability Assessment

To estimate whether the warrants generated by the LLM are acceptable, we employed a binary classification model based on Llama [‡]. We trained this model using a dataset containing 450 warrants with human subjective acceptability annotations, which was publicly released by Gupta et al. (Gupta et al., 2024). During training, the inputs to Llama consisted of claims, evidence, warrants, and the target variable–human annotations of warrant acceptability. The model was fine-tuned with LoRA (Hu et al.) for binary classification. The data was split into 70% training and 30% evaluation sets. During training, the hyperparameters were: 5 epochs, batch size of 4, learning rate of 5e-5, maximum input sequence length of 2048 tokens, LoRA rank of 16, scaling coefficient of 32, and dropout rate of 0.05. The modules where LoRA was applied were q_proj and v_proj. The trained model achieved an accuracy of 91.80%, a recall of 83.58%, and an F1 score of 87.5% on the evaluation dataset. Therefore, the trained model can at least reliably judge the acceptability of warrants in a general domain.

C Training Setting for Regression Model

For implementing the evaluation metric, we trained a regression model using Llama, with the input being review texts and human subjective evaluations of Objectivity from RottenReview. The training data comprised 70% of the dataset, and the evaluation data 30%. During training, the hyperparameters were set as follows: 5

[†] openai/gpt-oss-120b

[‡] meta-llama/Meta-Llama-3-8B

Table 6: Comparison of claim and evidence pair extraction results between state-of-the-art models in related work and our model.

	Claim tagging			Evidence linking	
	Precision	Recall	F1	EM	F1
Baseline	52.00	59.77	55.61	64.31	82.07
Our	**61.36**	53.13	**56.91**	**66.68**	71.98

epochs, a learning rate of 5e-5, a batch size of 8, and a maximum input length of 8,096 tokens. Regarding LoRA, the rank was set to 16, the scaling coefficient to 32, and the dropout rate to 0.05, applied to the q_proj and v_proj modules.

D claim and evidence extract

To implement SubstanScore, we replicated the claim tagging and claim and evidence linking experiments conducted by Guo et al. (Guo et al., 2023) using LLMs, with modifications to the training conditions for this specific experiment.

Claim tagging involves a token classification task (Segal et al., 2020) that extracts multiple spans containing claims from a given review text. Linking evidence involves a Question Answering (QA) extraction task (Rajpurkar et al., 2016) that extracts corresponding evidence spans from claim-review pairs. In our study, we used the state-of-the-art BERT-based model, ModernBERT (Warner et al., 2025)[§], which supports a maximum input token length of 8,096 tokens, to replicate both claim tagging and evidence linking experiments.

The claim and evidence span in the scientific review texts for this experiment were cropped to remove punctuation and whitespace, including spaces, tabs, and newlines, from both ends of the spans. These spans were derived from claim and evidence annotations manually created in SubstanReview. The training dataset comprised 440 reviews, and the evaluation dataset comprised 110 reviews. We combined all training and evaluation data and conducted 5-fold cross-validation to evaluate the model's performance.

D.1 Claim Tagging

The hyperparameters used for training the claim tagging model are as follows: batch size of 8, 20 epochs, learning rate of 4e-5, weight decay of 0.10, maximum gradient norm of 1.0, hidden layer dropout rate of 0.1, and attention mechanism dropout rate of 0.1. In prior research, positive and negative claims were distinguished by applying BIO (Beginning, Inside, Outside) encoding, training the token classification model to identify five classes: B-claim_positive, I-claim_positive, B-claim_negative, I-claim_negative, and O. However, since SubstanScore does not differentiate between positive and negative claims, our study simplifies the task by combining both into a single "claim" category for tagging. Consequently, the classes for token classification are B-Claim, I-Claim, and O. The evaluation of claim tagging and the span annotations in subsequent tasks define a span as the sequence starting from a B-Claim token, continuing with consecutive I-Claim tokens, and ending when O tokens appear.

D.2 Evidence Linking

The training for the evidence linking task involves inputting the claim and review sentences $R = \{r_1, r_2, \ldots, r_{|R|}\}$ along with the claims $C = \{c_1, c_2, \ldots, c_{|C|}\}$ as a combined sequence structured as $\{[CLS]c_1c_2 \cdots c_{|C|}[SEP]r_1r_2 \cdots r_{|R|}\}$. This input format is used to extract the span of evidence corresponding to each claim. The hyperparameters during training are batch size of 4, 16 epochs, learning rate of 1e-5, mixed precision training enabled via FP16, with a hidden layer dropout rate of 0.2 and an attention mechanism dropout rate of 0.2.

claim and evidence extraction is evaluated using the same metrics as in prior research, with the results shown in Table 6. These results demonstrate that the pre-trained ModernBERT model achieves performance comparable to the state-of-the-art models reported in previous studies. Therefore, for claim and evidence extraction in our experiments, we utilize this trained ModernBERT. The maximum input token length is set to 8,096 tokens, and the model is trained on the 440 reviews from the SubstanReview training dataset.

[§] answerdotai/ModernBERT-large