When Evidence Conflicts: Reliability-aware Meta-review Generation

Xinzhe Wang[1], Fei Tao[2], Jiang Xie[1], Hong Yu[1], Ye Wang[1]*
[1] Chongqing University of Posts and Telecommunications, China
[2]NewsBreak, USA

Abstract

Generating coherent meta-reviews from multiple peer reviews is challenging when reviewer evidence conflicts and varies in reliability. Existing approaches typically formulate meta-review generation as a multi-document summarization task and aggregate reviewer feedback uniformly, making it difficult to determine which opinions should be prioritized under disagreement. In this paper, we study meta-review generation through reliability-aware evidence aggregation. Our framework first extracts aspect-level opinions from peer reviews and identifies conflicting evidence within each aspect. It then estimates opinion-level support and review-level quality to measure evidence reliability. Based on these signals, the framework assigns reliability-aware weights to reviewer feedback, enabling the generator to prioritize better-supported arguments while preserving diverse perspectives. Experiments demonstrate that our method consistently improves meta-review generation over strong baselines on both automatic and human evaluations, with clear gains in conflict recognition and resolution under high-conflict review scenarios. The code and implementation details are publicly available at https://github.com/Wangxz729/reliability-aware-meta-review.

1 Introduction

Synthesizing multiple peer reviews into a meta-review is an important component of academic peer review. With the development of large language models (LLMs), recent work has explored meta-review generation by summarizing reviewer feedback from multiple sources (Du et al., 2024; Zeng et al., 2025; Li et al., 2024; Hossain et al., 2025; Idahl and Ahmadi, 2025). Despite promising progress, this task remains challenging when reviewer evidence conflicts. Reviewers may provide divergent assessments of the same paper in terms of novelty, soundness, empirical validity, or overall significance. Uniform aggregation can blur key disagreements and weaken the evidential basis of the meta-review.

Reviewer disagreement is common in academic peer review and often reflects differences in expertise, interpretation, or evaluation criteria rather than simple annotation noise (Basile et al., 2021; Xu and Jurgens, 2026). Prior work on disagreement modeling shows that conflicting judgments can offer complementary perspectives (Chu et al., 2021; Basile et al., 2021; Leonardelli et al., 2023). Thus, meta-review generation should not simply collapse disagreement into a single summary. Instead, it should weigh conflicting evidence when forming the final meta-review.

A further challenge is that conflicting reviewer evidence is not equally reliable. Some comments provide concrete observations, detailed reasoning, and well-supported critiques, whereas others are vague, speculative, or weakly justified. Existing approaches uniformly aggregate reviewer comments, leading to difficulty in prioritizing opinions when evidence conflicts. Therefore, reviewer disagreement should not be resolved by majority sentiment or uniform summarization alone. The key question is which pieces of evidence deserve greater influence when forming the final meta-review. This view is related to argument quality assessment, which evaluates textual claims according to factors such as specificity, justification, and evidential support (Wachsmuth et al., 2017; Habernal and Gurevych, 2016a).

Existing approaches provide limited mechanisms for addressing this problem. Most meta-review generation methods formulate the task as multi-document summarization and mainly improve generation quality through prompting strategies or iterative refinement (Li et al., 2023; Zeng

*Corresponding author: wangye@cqupt.edu.cn

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 15708–15723
October 24-29, 2026 ©2026 Association for Computational Linguistics

et al., 2025; Hossain et al., 2025). Sentiment-based consolidation methods can expose coarse agreement and disagreement patterns, but they typically treat opinions with the same polarity similarly (Li et al., 2024). More broadly, prior work on opinion and aspect-based summarization has explored structured opinion extraction, viewpoint consolidation, and evidence aggregation, but does not explicitly model the reliability of competing positive and negative opinions within the same review aspect. As a result, existing approaches can identify or organize reviewer disagreement, but still lack an explicit mechanism for determining which conflicting evidence is better supported and should receive greater influence during generation.

To address this limitation, we investigate meta-review generation via reliability-aware evidence aggregation. Our framework first extracts aspect-level opinions from peer reviews and detects conflicting evidence per aspect. It then estimates opinion-level support and review-level quality with an LLM-based evaluator, following recent work on LLM-based evaluation (Zheng et al., 2023). Using these reliability signals, the framework assigns adaptive weights to reviewer evidence to prioritize better-supported arguments during aggregation. Weighted evidence is structured into aspect-level representations to guide the generator for final meta-review production.

We evaluate our framework on the ORSum benchmark. Results show consistent improvements over strong baselines in both automatic and human evaluations, with larger gains on high-conflict cases. These findings highlight the value of reliability-aware aggregation when reviewer evidence conflicts.

Our contributions are summarized as follows:

- We formulate meta-review generation under reviewer disagreement as a reliability-aware evidence aggregation problem, where conflicting opinions are modeled at the aspect level rather than at the whole-review level.
- We introduce a two-level reliability-aware aggregation framework that combines opinion-level support with review-content quality and uses these signals to conditionally down-weight weak evidence while preserving competing viewpoints.
- We provide empirical evidence through component ablations, evaluator robustness analysis, controlled weight/order experiments, human evaluation, and conflict-focused LLM evaluation on the ORSum benchmark.

2 Related Work

2.1 Meta-review Generation

Meta-review generation aims to synthesize multiple reviewer comments and discussions into a coherent meta-review for academic decision-making. Early studies formulate this task as a multi-document summarization problem over peer reviews and author-reviewer discussions (Li et al., 2023; Zeng et al., 2025). Related work on multi-document summarization has also explored how to aggregate information from multiple source documents into coherent summaries (Fabbri et al., 2019).

More recently, meta-review generation has been studied in LLM-driven settings, where prompting strategies and multi-stage generation pipelines are commonly adopted. For example, CGI^2 decomposes meta-review generation into multiple steps with iterative self-refinement, while TELeR studies structured prompting strategies for assisting meta-review writing (Zeng et al., 2025; Hossain et al., 2025). Other studies further examine the potential and risks of using LLMs in scholarly peer review and meta-reviewing (Du et al., 2024; Ye et al., 2024). Although these methods improve generation fluency and coverage, they largely treat reviewer evidence as uniformly reliable, leaving the problem of conflicting evidence underexplored.

2.2 Evidence Aggregation in Opinion Summarization

Meta-review generation can be viewed as an opinion aggregation problem, where reviewer comments provide evaluative evidence across different aspects. Prior work on opinion and argument-aware summarization has explored aspect extraction, sentiment analysis, viewpoint consolidation, and evidence-aware aggregation (Angelidis et al., 2021; Wachsmuth et al., 2017; Habernal and Gurevych, 2016a; Basile et al., 2021). Recent studies further incorporate argumentation schemes, aspect key-point analysis, and decomposed aspect-aware modules for structured opinion summarization (Zhou et al., 2025; Tang et al., 2024; Li et al., 2025). However, scientific peer reviews pose a distinct challenge: opinions are aspect-dependent and argument-based, and positive and negative evaluations may directly conflict.

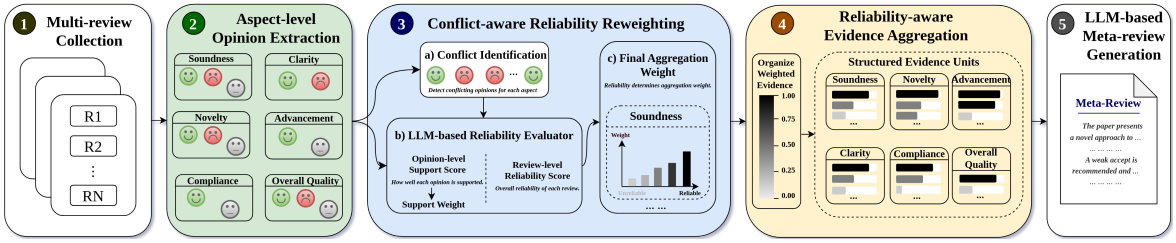

Figure 1: Overview of the proposed framework. The framework extracts aspect-level opinions from peer reviews, identifies conflicts within each aspect, estimates opinion-level support and review-content quality, performs conflict-conditioned reliability-aware evidence aggregation, and generates the final meta-review from the resulting weighted evidence.

In peer-review summarization, recent sentiment-based consolidation methods use polarity signals to guide meta-review generation (Li et al., 2024). While such approaches make agreement and disagreement more explicit, sentiment alone does not indicate the quality or evidential strength of an opinion. Opinions with the same polarity may differ substantially in specificity, justification, and support. Our approach therefore focuses on estimating the relative reliability of competing opinions and incorporating both opinion-level support and review-level reliability into conflict-aware evidence aggregation.

2.3 Reliability of Conflicting Evidence

Conflicting evidence is common in subjective evaluation settings, where different judgments may reflect diverse perspectives rather than noise. Prior work on disagreement modeling shows that such disagreement can contain meaningful signals and should not always be reduced to a single consensus label (Basile et al., 2021; Leonardelli et al., 2023; Xu and Jurgens, 2026). This is especially relevant to peer review, where reviewer disagreement may arise from differences in expertise, criteria, or interpretation.

Reliability has also been studied in crowdsourced learning and argument quality assessment. Crowdsourced learning estimates annotator reliability to reduce noisy supervision (Chu et al., 2021), while argument quality assessment evaluates textual claims through specificity, justification, plausibility, and evidential support (Wachsmuth et al., 2017; Habernal and Gurevych, 2016a). However, these reliability signals have not been fully used in meta-review generation under conflicting reviewer evidence. Our work addresses this gap by estimating opinion-level support and review-level quality, weighting reviewer evidence during generation.

3 Method

We propose a reliability-aware evidence aggregation framework for meta-review generation under reviewer disagreement. Instead of uniformly aggregating reviewer comments, our framework explicitly models the reliability of different opinions and prioritizes better-supported evidence during generation.

Figure 1 presents an overview of the proposed reliability-aware evidence aggregation framework. Starting from peer reviews, the framework extracts aspect-level opinions, identifies reviewer disagreement, estimates opinion reliability, and organizes reviewer evidence for meta-review generation.

Given a set of peer reviews for a paper, our goal is to generate a coherent meta-review that appropriately resolves conflicting reviewer opinions while preserving important supporting evidence.

In this paper, we distinguish between a *review*, which refers to an entire reviewer report, and an *opinion*, which denotes an aspect-level evaluative statement extracted from a review.

3.1 Opinion Extraction and Conflict Characteristics

We begin by extracting structured opinion units from raw peer reviews and organizing them by predefined review aspects, such as soundness, clarity, novelty, and overall quality. We perform aspect-level opinion extraction using an LLM-based structured extraction prompt. For each review, the model identifies aspect labels, opinion, and sentiment polarity labels (positive, neutral, negative). Extracted opinions are then normalized into structured opinion units for subsequent conflict analysis and aggregation. To improve consistency, all extraction prompts use deterministic decoding (temperature=0). Example extraction prompts are provided in Appendix F.

Aspect	Positive (%)	Neutral (%)	Negative (%)
Soundness	15.14	4.62	80.24
Clarity	31.96	6.75	61.29
Advancement	47.61	3.44	48.95
Novelty	59.52	4.56	35.92
Overall Quality	63.00	10.46	26.54
Compliance	11.51	32.85	55.64

Table 1: Sentiment distribution across review aspects (in percentage).

Aspect	Conflict Ratio (%)
Soundness	42.05
Clarity	59.56
Advancement	64.52
Novelty	53.94
Overall Quality	40.44
Compliance	7.95

Table 2: Proportion of papers with conflicting opinions per aspect.

To better understand the nature of reviewer disagreement, we first analyze the distribution of sentiment across different aspects. Table 1 presents the normalized sentiment distribution aggregated over the dataset.

As shown in Table 1, different aspects exhibit substantially different sentiment patterns. For example, *soundness* is dominated by negative opinions, indicating that reviewers tend to be more critical of methodological rigor. In contrast, *novelty* and *overall quality* receive relatively more positive feedback. Notably, aspects such as *advancement* present a nearly balanced distribution between positive and negative sentiments, suggesting a higher likelihood of disagreement. This variation indicates that disagreement is not uniformly distributed, but depends on the nature of the evaluation aspect.

To further quantify the prevalence of disagreement, we compute the proportion of papers for which conflicting sentiments exist within each aspect. An aspect is considered conflicting if both positive and negative opinions are present. The results are shown in Table 2.

The results reveal that reviewer disagreement is both pervasive and highly aspect-dependent. In particular, aspects such as *advancement*, *clarity*, and *novelty* exhibit high conflict ratios, indicating that reviewers frequently hold opposing views on these dimensions. In contrast, *compliance* shows relatively low disagreement, suggesting more consistent and objective judgments.

To further characterize disagreement at the paper level, we define the conflict ratio of a paper as the proportion of aspects containing both positive and negative opinions among all extracted aspects:

$$\text{ConflictRatio}(p) = \frac{\sum_{a \in \mathcal{A}_p} C_a}{|\mathcal{A}_p|}, \tag{1}$$

where $\mathcal{A}_p$ denotes the set of extracted aspects for paper p, and $C_a \in \{0, 1\}$ indicates whether aspect a contains conflicting opinions as defined in Equation 3.

These observations highlight two key challenges for meta-review generation. First, conflicts are widespread rather than exceptional, making naive aggregation strategies insufficient. Second, the degree of disagreement varies across aspects, implying that conflict handling should be adaptive rather than uniform. Moreover, aspects with higher conflict ratios tend to involve more subjective evaluation criteria, which further complicates reliable aggregation. These findings motivate the need for a conflict-aware mechanism that can explicitly account for disagreement and differentiate the influence of competing opinions.

3.2 Conflict-aware Opinion Support and Down-weighting

To resolve reviewer disagreement in a principled manner, we propose a conflict-aware support estimation and down-weighting mechanism that assigns different importance to opinions based on their strength.

Conflict-aware Setting For each aspect a, we denote its associated opinion set as

$$\mathcal{O}_a = \{o_1, o_2, \ldots, o_k\}. \tag{2}$$

Each opinion $o \in \mathcal{O}_a$ is associated with a sentiment label $s(o) \in \{\text{pos}, \text{neg}, \text{neu}\}$. We define an aspect as conflicting only when both positive and negative opinions coexist under the same aspect:

$$C_a = \mathbb{I}\left[\exists o_i, o_j \in \mathcal{O}_a : s(o_i) = \text{pos}, s(o_j) = \text{neg}\right] \tag{3}$$

For non-conflicting aspects, we retain all extracted opinions and apply review-level reliability through q_r. For conflicting aspects, we additionally apply opinion-level support weighting to differentiate the relative influence of competing positive and negative evidence.

Opinion Support Estimation To quantify the strength of each opinion, we evaluate four factors: *specificity*, *rationality*, *justification*, and *academic plausibility*. These dimensions are adapted from prior work on argument quality and convincingness (Wachsmuth et al., 2017; Habernal and Gurevych, 2016a,b). Each factor f_i is rated on a discrete scale from 1 to 5 and normalized to the range $[0, 1]$:

$$\tilde{f}_i = \frac{f_i - 1}{4}. \tag{4}$$

The overall support score is computed as a weighted sum:

$$\text{support}(o) = \sum_{i=1}^{4} \alpha_i \tilde{f}_i(o), \tag{5}$$

where $\sum_{i=1}^{4} \alpha_i = 1$. In our implementation, we adopt uniform weights $\alpha_i = 0.25$ to avoid introducing additional task-specific hyperparameters and to prevent the scoring function from implicitly favoring one argument-quality dimension without sufficient empirical justification. We further examine the sensitivity of this choice in Appendix A. The support score is estimated using an LLM-based evaluator according to predefined argument-quality criteria. Detailed prompting and scoring settings are described in Section 4.

Down-weighting Mechanism Given a threshold $\tau \in [0, 1]$, for opinions satisfying $\text{support}(o) < \tau$, we compute a normalized support deficiency:

$$\delta(o) = \frac{\tau - \text{support}(o)}{\max_{o' \in \mathcal{O}_a} (\tau - \text{support}(o'))}. \tag{6}$$

We adopt a smooth reliability-aware gating function to determine the contribution of each opinion during aggregation.The purpose of this function is to reduce the influence of weakly supported evidence without completely discarding minority opinions, which may still contain important concerns. The resulting weight assigned to an opinion, denoted as the *support weight*, is defined as:

$$w(o) = \begin{cases} 1, & \text{support}(o) \geq \tau, \\ \sigma\big(\lambda(1 - \delta(o))\big), & \text{support}(o) < \tau. \end{cases} \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and λ controls the sharpness of the weighting transition.

This formulation provides a smooth reliability-aware weighting strategy that progressively down-weights weakly supported opinions instead of discarding them entirely. We do not claim that the sigmoid form is theoretically optimal; rather, it is an empirically effective implementation of conflict-conditioned reliability weighting.

Discussion The proposed mechanism provides a smooth and continuous weighting strategy. Instead of discarding low-support opinions, it progressively reduces their influence while preserving diverse viewpoints. This design enables more robust and fine-grained resolution of conflicting reviewer opinions.

3.3 Review-level Reliability Modeling

Beyond opinion-level support estimation, we further model review-level reliability to capture variation in review content quality across different reviews. The intuition is that some review texts provide clearer, more specific, and better-justified evidence than others. Ignoring such variation may lead to over-reliance on weakly supported review content even when individual opinions appear reasonable.

Importantly, our framework does not attempt to estimate the intrinsic competence or long-term credibility of individual reviewers. Instead, reliability estimation is applied only to the textual quality of review content within the current paper context.

To this end, we estimate a review-level reliability score based on several textual quality indicators, including clarity, specificity, constructiveness, and justification strength. Each review is associated with a reliability score $q_r \in [0, 1]$, reflecting the estimated reliability of the review content for the current paper. Review-level reliability scores are also computed using an independent LLM-based evaluator following predefined quality rubrics.

Formally, given an opinion o extracted from review r, we incorporate review-level reliability by modulating its opinion-level support weight. The resulting *final aggregation weight* is defined as:

$$\tilde{w}(o) = w(o) \cdot q_r, \tag{8}$$

where $w(o)$ denotes the opinion-level support weight defined in Section 3.2, and q_r represents the reliability score of review r.

This formulation ensures that opinions extracted from higher-quality review content contribute more

Method	Backbone Generator: Qwen3				Backbone Generator: DeepSeek			
	R-1	R-2	R-L	BERTScore	R-1	R-2	R-L	BERTScore
CGI^2 (24' IJCAI-W)	26.91	3.55	14.42	83.55	27.38	4.85	15.27	83.39
TELeR Prompt level 3 (25' NAACL)	23.97	4.20	12.20	81.31	23.53	3.95	12.11	80.94
TELeR Prompt level 4 (25' NAACL)	21.12	**4.34**	11.01	81.28	21.29	4.50	11.11	81.35
Sentiment Consolidation (24' ACL)	22.57	4.28	11.33	82.00	24.78	4.75	13.07	82.61
Ours	**27.90**†	4.17	**14.54**	**83.90**†	**28.58**†	**5.01**	**15.29**	**83.96**†

Table 3: Main results on the ORSum benchmark. Our framework consistently improves ROUGE-1, ROUGE-L, and BERTScore across both backbone generators, while ROUGE-2 shows mixed results. † indicates statistically significant improvement over CGI^2 under the paired bootstrap test.

strongly during aggregation, while still preserving diverse reviewer viewpoints. By combining opinion-level support estimation with review-level reliability modeling, the framework achieves more robust evidence aggregation under reviewer disagreement.

3.4 Reliability-aware Evidence Aggregation

Given a set of weighted opinions, we organize them into structured evidence units for meta-review generation. Each evidence unit contains the aspect label, opinion, sentiment polarity, and the final aggregation weight.

For each aspect a, we construct a ranked evidence set:

$$E_a = \text{Sort}\left(\{(o, \tilde{w}(o))\}_{o \in \mathcal{O}_a}\right), \tag{9}$$

where opinions are ordered by their final aggregation weights while the weights themselves are explicitly provided to the generator. The ordering is therefore used only to organize the evidence; the numerical weights constitute the primary reliability signal.

The final meta-review is generated by conditioning the language model on the aggregated evidence across all aspects:

$$M = G(\{E_a\}_{a \in \mathcal{A}}), \tag{10}$$

where $G(\cdot)$ denotes the generator language model.

This generation method organizes evidence according to both aspect structure and estimated reliability. By explicitly exposing the numerical weights to the generator, the framework allows the generator to distinguish the relative influence of competing evidence rather than relying solely on their input order.

4 Experiments

4.1 Experimental Setup

Dataset and Baselines. We evaluate our framework on the ORSum benchmark (Zeng et al., 2025), a widely used dataset for meta-review generation. ORSum contains peer reviews and corresponding meta-reviews written by area chairs across multiple scientific venues, making it particularly suitable for studying reviewer disagreement and opinion aggregation.

We compare our method with several representative baselines, including CGI^2 (Zeng et al., 2025), TELeR (Hossain et al., 2025), and the Sentiment Consolidation Framework (Li et al., 2024).

Models and Evaluation Metrics. We conduct experiments using Qwen3 and DeepSeek as backbone generation models responsible for meta-review generation, while GPT-4o-mini is used separately as the evaluator model for reliability estimation. To assess the robustness of reliability estimation, we additionally evaluate GPT-4o-mini, Qwen3-32B, and DeepSeek-V3 as alternative evaluator models under a controlled setting.

Following prior work on meta-review generation and summarization, we adopt ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore (F1) as automatic evaluation metrics. ROUGE measures lexical overlap between generated and reference meta-reviews, while BERTScore evaluates semantic similarity at the contextual embedding level.

To further analyze performance under reviewer disagreement, we additionally partition the test set into low-conflict, medium-conflict, and high-conflict subsets according to aspect-level conflict ratios.

Implementation Details. For reliability estimation, all opinion support scores and review-level reliability scores are computed using an LLM-based

evaluator under deterministic decoding settings (temperature=0). The reliability threshold τ is set to 0.5 and the sigmoid sharpness parameter λ is set to 5 in all experiments. Implementation details and hyperparameter settings are provided in Appendix B.

Reliability Estimation Details. We use GPT-4o-mini as the evaluator model for both opinion support estimation and review-level reliability scoring. The evaluator is independent from the backbone generation models to reduce evaluation-generation coupling.

For each opinion, the evaluator scores four dimensions: specificity, rationality, justification strength, and academic plausibility. Each dimension is rated on a discrete 1–5 scale according to predefined rubrics. The final support score is computed as the normalized average across all dimensions.

For review-level reliability estimation, the evaluator assesses the overall quality of review content based on clarity, constructiveness, specificity, and evidential grounding. All evaluations are performed under deterministic decoding settings with temperature=0.

Complete scoring rubrics and prompts are provided in Appendix C and Appendix F.

Reliability Score Validation We randomly sample 200 opinions and have two annotators score opinion quality on specificity, rationality, justification, and academic plausibility. The estimated support scores correlate well with human judgments: Spearman 0.71, Pearson 0.68, and inter-annotator Cohen's κ 0.74. This confirms that the LLM-based evaluator captures meaningful signals aligned with human perception of argument quality.

4.2 Main Results

Table 3 compares performance across backbone models. Our framework consistently improves ROUGE-1 and BERTScore over the strongest baseline under both Qwen3 and DeepSeek. Compared with CGI^2, ROUGE-1 improves from 26.91 to 27.90 under Qwen3 and from 27.38 to 28.58 under DeepSeek, while BERTScore improves by 0.35 and 0.57 points, respectively. ROUGE-2 results are mixed under Qwen3, whereas our method achieves the highest ROUGE-2 under DeepSeek. We therefore avoid claiming uniform improvements across all automatic metrics.

Compared with CGI^2, which uses iterative refinement, our method further improves ROUGE and BERTScore by 1.5–2.0%, indicating complementary gains from reliability-aware aggregation. TELeR, relying on prompts but ignoring evidence strength, performs worse on most metrics. Sentiment Consolidation treats opinions with the same polarity equally, while our framework prioritizes better-supported opinions, yielding more reliable meta-reviews.

Results are consistent across Qwen3 and DeepSeek, showing model-agnostic robustness. Paired bootstrap tests confirm statistically significant improvements on ROUGE-1 and BERTScore. Overall, the gains arise from reliability-aware evidence aggregation rather than generation variance.

Statistical significance. We perform 500 paired bootstrap resamples over the full ORSum test set. Compared with CGI^2, the strongest baseline on ROUGE-1 and BERTScore, the improvements are statistically significant on ROUGE-1 ($p < 0.01$) and BERTScore ($p < 0.05$).

Evaluator LLM	R-1	BERTScore
GPT-4o-mini	29.15	83.84
Qwen3-32B	28.96	83.81
DeepSeek-V3	29.01	83.83

Table 4: Robustness to the choice of reliability evaluator on 100 randomly sampled test papers. Downstream performance varies only slightly across evaluator models.

4.3 Evaluator Robustness

To examine whether the effectiveness of reliability estimation depends on a particular evaluator model, we conduct a controlled sensitivity analysis on 100 randomly sampled test papers. We use GPT-4o-mini, Qwen3-32B, and DeepSeek-V3 as reliability evaluators while keeping the generator, prompts, sampled papers, and decoding settings fixed. Only the evaluator model is changed.

The maximum downstream variation is 0.19 ROUGE-1 and 0.03 BERTScore. The evaluator models also show strong agreement in opinion-level reliability estimation, with pairwise Spearman correlations of 0.82–0.87 and Kendall's τ values of 0.71–0.76. These results suggest that the observed performance is not driven solely by one particular evaluator model.

4.4 Performance Across Conflict Levels

To assess our framework under reviewer disagreement, we evaluate methods on low-, medium-, and

Method	Low (N=167)	Med (N=166)	High (N=167)
CGI^2 (24' IJCAI-W)	26.88	26.34	25.71
TELeR L3 (25' NAACL)	23.51	22.98	22.83
TELeR L4 (25' NAACL)	21.92	21.53	20.94
Sent. Consol. (24' ACL)	22.93	22.51	21.24
Ours	**27.82**	**27.83**	**28.02**

Table 5: ROUGE-1 results across different conflict levels. Our method achieves more prominent improvements in high-conflict scenarios, verifying the efficacy of reliability-aware aggregation for resolving reviewer disagreements.

high-conflict subsets. Each subset contains 167, 166, and 167 papers, respectively.

Table 5 shows that the performance gains of our method increase with conflict level. On the high-conflict subset, our framework achieves ROUGE-1 of 28.02, outperforming CGI^2, which scores 25.71, and TELeR L4, which scores 20.94. The absolute ROUGE-1 scores are also higher in high-conflict cases, likely due to richer evaluative evidence. The key observation is the substantially larger relative gains achieved by our framework under severe disagreement.

On the low-conflict subset, our framework achieves ROUGE-1 of 27.82, compared to 26.88 for CGI^2, showing stable improvements even when reviewer agreement is high. These results support our central hypothesis that explicitly modeling evidence reliability is crucial for resolving conflicting reviewer opinions.

4.5 Ablation Study

We conduct ablation studies to evaluate the contribution of each component of our framework (Table 6): **Uniform** assigns equal weights to all opinions; **Support-only** uses only opinion-level support; **Reliability-only** uses only review-level reliability; **No Conflict Modeling** ranks opinions by reliability but ignores explicit conflict-aware grouping; **Full Model** denotes the complete framework.

Removing review-level reliability reduces ROUGE-1 from 27.90 to 27.35 and BERTScore from 83.90 to 83.61, while removing opinion-level support drops ROUGE-1 to 23.08 and BERTScore to 81.97. Eliminating conflict-aware aggregation slightly reduces ROUGE-1 to 24.32 and BERTScore to 82.38, especially on high-conflict samples. These results indicate that opinion-level support contributes most, but all components are complementary; the full model consistently achieves the best performance.

Variant	R-1	BERTScore
Uniform	22.71	81.22
Support-only	27.35	83.61
Reliability-only	23.08	81.97
No Conflict Modeling	24.32	82.38
Full Model	**27.90**	**83.90**

Table 6: Ablation study of different framework components. Both opinion-level support estimation and review-level reliability contribute substantially to overall performance.

4.6 Human Evaluation

Automatic metrics may not fully capture a system's ability to recognize and resolve reviewer disagreement. We therefore conduct human evaluation focusing on conflict handling quality.

We randomly sample 104 papers from the test set and evaluate generated meta-reviews through human assessment. Two independent annotators conduct blind evaluation using a 1–5 Likert scale across the following dimensions (see Appendix D for full human evaluation protocol).:

- **Conflict Awareness**: whether the generated meta-review correctly recognizes reviewer disagreement.
- **Resolution Quality**: whether conflicting opinions are resolved coherently and reasonably.
- **Evidence Grounding**: whether the generated conclusions are supported by reviewer evidence.
- **Overall Coherence**: the overall readability, fluency, and consistency of the generated meta-review.

The average Cohen's κ between annotators is 0.72, indicating substantial agreement.

As shown in Table 7, our framework achieves the highest ratings across all dimensions. The largest improvements appear in Conflict Awareness and Resolution Quality, confirming the effectiveness of reliability-aware evidence aggregation under reviewer disagreement. Compared with sentiment-based aggregation, our method produces meta-reviews that are better grounded and less prone to overemphasizing unsupported opinions.

4.7 Conflict-focused LLM Evaluation

Automatic reference-based metrics may not fully capture whether a generated meta-review recognizes disagreement, resolves competing opinions,

Method	Conflict Awareness	Resolution Quality	Evidence Grounding	Coherence
CGI^2 (24' IJCAI-W)	2.81	2.74	3.69	3.92
TELeR Prompt level 3 (25' NAACL)	2.57	2.78	3.41	3.71
TELeR Prompt level 4 (25' NAACL)	2.43	2.51	3.68	3.62
Sentiment Consolidation (24' ACL)	3.64	3.08	3.79	3.65
Ours	**4.56**	**4.21**	**4.18**	**3.94**

Table 7: Human evaluation results on a 1–5 Likert scale. Our framework achieves the highest ratings across all dimensions, particularly on conflict awareness and resolution quality.

Method	Conflict	Resolution	Evidence
CGI^2	3.02	2.54	**3.92**
TELeR (Level 3)	2.23	2.57	3.56
TELeR (Level 4)	2.45	2.48	3.62
Senti. Consol.	3.56	3.15	3.68
Ours	**4.31**	**3.83**	3.85

Table 8: GPT-4o-based conflict-focused evaluation on the same 104 papers used for human evaluation. The judge assesses the generated meta-review against the source reviews without access to the reference meta-review.

and grounds its conclusions in the source reviews. We therefore conduct a complementary LLM-as-a-judge evaluation using GPT-4o on the same 104 papers used for human evaluation.

GPT-4o is given the source reviews and one generated meta-review, with system identities and the reference meta-review hidden. The judge evaluates each output independently using a fixed 1–5 rubric covering Conflict Awareness, Resolution Quality, and Evidence Grounding. We use deterministic decoding and the same evaluation prompt for all systems.

As shown in Table 8, our framework achieves the highest scores on all three dimensions, with the largest gains in Conflict Awareness and Resolution Quality. This suggests that its improvements extend beyond reference-based similarity metrics. A concrete case study illustrating conflict-aware aggregation is presented in Appendix E.

5 Conclusion

Meta-review generation is challenging when reviewer opinions conflict, as uniform aggregation can obscure important disagreements and weaken evidence. To address this, we propose a reliability-aware evidence aggregation framework that explicitly evaluates the quality of each opinion and review. Our method combines opinion-level argument support, review-level quality signals, and conflict-aware weighting to prioritize better-supported evidence during meta-review generation. Experiments on the ORSum benchmark demonstrate consistent improvements over strong baselines in both automatic metrics and human evaluations, with particularly larger gains on high-conflict cases. These results provide empirical evidence that explicitly modeling evidence reliability can improve meta-review generation.

More broadly, our work highlights the potential of reliability-aware evidence aggregation for scientific opinion summarization and provides a basis for further study of trustworthy, LLM-assisted peer review systems.

Limitations

While our framework improves meta-review generation under reviewer disagreement, several limitations remain. First, reliability estimation depends on LLM-based evaluation, which may introduce biases or inconsistencies. Second, our evaluation is conducted on the ORSum benchmark alone, cross-dataset validation is an important direction for future work. Finally, minority opinions may sometimes highlight important flaws, and reliability-aware aggregation could underemphasize these dissenting viewpoints.

Future work could explore confidence-aware aggregation, richer conflict modeling, and human-in-the-loop meta-review generation.

Ethical Considerations

Our framework is intended to assist, not replace, human reviewers. Reliability scores assess review content rather than reviewer competence, and final decisions should remain with humans.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62306056, 62136002, and 62221005).

References

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Zhendong Chu, Jing Ma, and Hongning Wang. 2021. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5832–5840.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

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. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7763–7803, Albuquerque, New Mexico. Association for Computational Linguistics.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

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

Miao Li, Jey Han Lau, and Eduard Hovy. 2024. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177, Bangkok, Thailand. Association for Computational Linguistics.

Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Decomposed opinion summarization with verified aspect-aware modules. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24805–24841, Vienna, Austria. Association for Computational Linguistics.

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Erik Cambria. 2024. Prompted aspect key point analysis for quantitative review summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in nlp. *Preprint*, arXiv:2601.09065.

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 preprint arXiv:2412.01708*.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2025. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *Artificial Intelligence for Research and Democracy*, pages 20–38, Singapore. Springer Nature Singapore.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Wendi Zhou, Ameer Saadat-Yazdi, and Nadin Kökciyan. 2025. Aspect-based opinion summarization with argumentation schemes. In *Proceedings of the 12th Argument mining Workshop*, pages 116–125, Vienna, Austria. Association for Computational Linguistics.

A Weight and Ordering Analysis

A.1 Controlled analysis of numerical weights and evidence ordering.

To distinguish the effect of explicit numerical weights from the effect of evidence ordering, we conduct a controlled experiment on 100 randomly sampled test papers. All settings use the same sampled papers, generator, prompts, and decoding configuration; only the numerical-weight information and evidence order are changed.

Under random ordering, adding explicit numerical weights improves ROUGE-1 by 1.12 and BERTScore by 0.24. With numerical weights retained, sorting the evidence adds 0.21 ROUGE-1 and 0.02 BERTScore. These results suggest that numerical weights provide the primary contribution, whereas evidence ordering has a smaller complementary effect. Thus, our approach does not rely solely on placing high-weight evidence earlier in the input.

Setting	ROUGE-1	BERTScore
Weight + descending order	29.15	83.84
Weight + random order	28.94	83.82
No weight + random order	27.82	83.58

Table 9: Controlled analysis of numerical weights and evidence ordering. Explicit numerical weights provide the primary gain, while ordering provides a smaller complementary effect.

A.2 Sensitivity to Support-Factor Weights

We further examine whether the equal weighting of the four support dimensions materially affects performance. We compare the default equal weighting with moderate changes that emphasize justification or specificity.

The maximum variation is 0.09 ROUGE-1 and 0.03 BERTScore, indicating limited sensitivity to moderate changes in factor weighting. This supports our use of equal weights as a simple setting that avoids introducing additional task-specific hyperparameters.

Weighting Scheme	ROUGE-1	BERTScore
Equal $(1, 1, 1, 1)$	29.15	83.84
Justification-heavy $(1, 1, 2, 1)$	29.17	83.84
Specificity-heavy $(2, 1, 1, 1)$	29.08	83.81

Table 10: Sensitivity to the relative weighting of the four opinion-support dimensions. Moderate changes produce only small performance differences.

B Hyperparameter Settings

Table 11 summarizes the main hyperparameter settings used in our experiments.

Parameter	Value
Support threshold τ	0.5
Evaluator temperature	0
Generator temperature	0.7
Max generation length	512
Top-p	0.9
λ	5

Table 11: Main hyperparameter settings.

C Reliability Scoring Rubric

Table 12 summarizes the scoring criteria used for opinion-level support factor evaluation. The rubric is designed to encourage discriminative scoring across opinions with different levels of evidence quality. Higher scores indicate that an opinion is more concrete, better justified, and more academically reliable, thereby receiving a larger support weight during aggregation.

Score	Specificity	Rationality	Justification	Academic Plausibility
1	Vague or generic	Illogical or emotional	No supporting evidence	Academically weak
3	Partially concrete	Reasonably logical	Some supporting rationale	Reasonable judgment
5	Highly concrete	Well-reasoned and consistent	Strong supporting evidence	Academically convincing

Table 12: Scoring rubric for opinion-level support factor evaluation.

Table 13 presents the review-level quality dimensions used for reviewer reliability estimation. These dimensions evaluate the overall quality of reviewer comments from multiple perspectives, including clarity, specificity, constructiveness, and professionalism.

Dimension	Description
Clarity	Whether the review comments are easy to understand and clearly written.
Specificity	Whether the review discusses concrete aspects of the paper.
Constructiveness	Whether the review provides actionable suggestions or useful feedback.
Professionalism	Whether the review maintains an objective and professional tone.

Table 13: Review-level quality dimensions used for reliability estimation.

D Human Evaluation Protocol

Both annotators were graduate NLP researchers who had previously served as reviewers for peer-reviewed NLP/AI conferences or journals. The evaluation was conducted independently and blindly, with system identities hidden from the annotators. For each paper, the final score for each dimension was obtained by averaging the two annotators' ratings without an additional adjudication step. The mean Cohen's κ across the four evaluation dimensions was 0.72.

E Case Study

To further illustrate the effectiveness of the proposed reliability-aware aggregation framework, we present an additional case study from the ORSum benchmark.

Table 14 shows an example where reviewers express conflicting opinions regarding the novelty of the submission. While multiple reviewers question the originality of the proposed method, other reviewers emphasize the empirical effectiveness and robustness improvements. Our framework successfully identifies the disagreement and generates a balanced meta-review that reflects both supportive and critical viewpoints.

Aspect	Novelty
R1	"...The paper proposes a method for learning robust binary neural networks from random initialization. ..." (Positive)
R2	"...The paper's proposed three techniques are not systematical but auxiliary model compression strategies. ..." (Negative)
R3	"...The novelty of this work is a bit limited, as it combines existing techniques. ..." (Negative)
Meta-review	"...The paper studies an interesting problem and demonstrates strong empirical performance, although reviewers raised concerns regarding the methodological novelty of combining existing techniques. ..."

Table 14: Case study illustrating conflict-aware aggregation under reviewer disagreement.

F Prompt Templates

Our framework employs four prompt-based modules: aspect-level opinion extraction (Table 15), opinion-level support factor evaluation (Table 16), review-level quality evaluation (Table 17), and reliability-aware meta-review generation (Table 18). For reproducibility, we present the corresponding prompt templates used in each stage of the framework.

You are an expert in extracting and structuring original sentences from academic paper reviews.

Given a paper and its multiple reviews, perform the following tasks.

Objectives:
(1) Extract clearly expressed opinion sentences from each individual review.
(2) Do not merge, deduplicate, summarize, or paraphrase opinions.
(3) Group the opinions according to review aspects.
(4) Annotate each opinion with its sentiment type.
(5) Strictly preserve the semantics of the original review text.

Core Constraints:
(1) The opinion content must come directly from the review text.
(2) Prefer sentence-level extraction.
(3) Only minor formatting modifications are allowed.
(4) Do not explain the reviewer's motivation.
(5) Do not rewrite sentences as "the reviewer thinks/notes/says".
(6) Do not merge multiple sentences into a new opinion.
(7) Do not introduce information not explicitly present in the review.

Review Aspects (STRICT):
Each opinion must belong to one of the following six aspects only.
(1) novelty: Whether the paper presents new ideas, methods, or perspectives.
(2) soundness: Whether the technical approach, methodology, experiments, or reasoning are correct and reliable.
(3) clarity: Whether the paper is clearly written and easy to understand.
(4) advancement: Whether the work advances the state of the art or contributes meaningful progress to the field.
(5) compliance: Whether the paper follows conference requirements such as reproducibility, ethical standards, citations, or formatting.
(6) overall_quality: The reviewer's overall evaluation of the paper.

Sentiment Types:
(1) positive: explicitly praises or affirms the paper.
(2) negative: explicitly criticizes or expresses dissatisfaction.
(3) neutral: objective description with no clear evaluative polarity.

Special Rule:
If a sentence contains both positive and negative opinions, split it into two separate opinions whenever possible.

Output Format:
The output must strictly follow JSON format.

```
{
"paper_id": "<paper ID>",
"aspects": {
  "<aspect_name>": [
    {
      "review": "<reviewer ID>",
      "opinion": "<original opinion sentence>",
      "sentiment": "<positive | negative | neutral>"
    }
  ]
}
}
```

Important Constraints:
(1) Output JSON only.
(2) Do not output explanations.
(3) Ensure the JSON can be parsed directly.
(4) Do not modify the original meaning of the sentence in `opinion`.

Table 15: Prompt template for aspect-level opinion extraction.

You are an expert in academic peer review analysis.

Given opinion sentences under the same review aspect, evaluate four support factors for each opinion.

Each factor should be rated on a scale from 1 to 5:
(1) 1 = very weak / not satisfied at all
(2) 2 = weak
(3) 3 = moderate
(4) 4 = strong
(5) 5 = very strong

Important Constraints:
(1) Use the full range (1–5); do not collapse everything to 3 or 4.
(2) Be discriminative: different opinions should receive different scores whenever appropriate.
(3) Avoid assigning identical scores unless the opinions are truly similar in quality.

Factor Definitions:
(1) Specificity: Does the opinion refer to concrete components of the paper, such as methods, experiments, datasets, or results?
(2) Rationality: Is the opinion logically sound and non-emotional?
(3) Justification: Does the opinion provide reasoning, explanation, or supporting evidence?
(4) Academic Plausibility: Is the opinion a plausible and meaningful academic judgment?

Output Format:
The output must strictly follow JSON format.

```
{
“opinions”: [
  {
    “review”: ”<reviewer ID>”,
    “opinion”: ””,
    “sentiment”: ””,
    “f1_specificity”: 1,
    “f2_rationality”: 1,
    “f3_justification”: 1,
    “f4_academic_plausibility”: 1
  }
]
}
```

Important Constraints:
(1) Output JSON only.
(2) Do not output explanations.
(3) Ensure the JSON can be parsed directly.
(4) Assign integer scores only.

Table 16: Prompt template for opinion-level support factor evaluation.

You are an expert in evaluating the quality of academic peer reviews.

You will be given all opinion sentences written by a single reviewer for a specific paper. These sentences have already been extracted from the review text and classified by review aspect and sentiment.

Your task is not to judge whether the reviewer is correct. Instead, evaluate the overall review quality across four dimensions.

Evaluation Dimensions:
(1) clarity: Are the comments easy to understand and clearly written?
(2) specificity: Do the comments reference concrete aspects of the paper?
(3) constructiveness: Do the comments provide useful suggestions or actionable feedback?
(4) professionalism: Are the comments respectful, objective, and professional?

Scoring Rules:
Each dimension must be scored from 1 to 5.
(1) 1 = very poor
(2) 2 = weak
(3) 3 = moderate
(4) 4 = good
(5) 5 = excellent

Important Constraints:
(1) Each opinion is associated with a weight between 0 and 1, indicating its reliability.
(2) Opinions with lower weights are less trustworthy and should have less influence on the final judgment.
(3) Focus more on high-weight opinions when evaluating overall review quality.
(4) Provide concise reasons for each quality score.

Output Format:
The output must strictly follow JSON format.

```
{
“review”: "<reviewer ID>",
“quality_scores”: {
  “clarity”: {
    “score”: 1,
    “reason”: "..."
  },
  “specificity”: {
    “score”: 1,
    “reason”: "..."
  },
  “constructiveness”: {
    “score”: 1,
    “reason”: "..."
  },
  “professionalism”: {
    “score”: 1,
    “reason”: "..."
  }
},
“overall_assessment”: "brief summary”
}
```

Important Constraints:
(1) Output JSON only.
(2) Do not output explanations outside the JSON object.
(3) Ensure the JSON can be parsed directly.
(4) All scores must be integers between 1 and 5.

Table 17: Prompt template for review-level quality evaluation.

You are an experienced Area Chair writing a meta-review for a top-tier academic conference.

You are given a structured evidence package extracted from peer reviews.

The following processing steps have already been applied:
(1) sentence-level opinion extraction
(2) aspect grouping
(3) conflict detection
(4) sentence-level reliability estimation
(5) reviewer-level reliability estimation

Each opinion have one reliability weight: final_weight

Evidence Interpretation Rules:
(1) Give higher attention to opinions with larger `final_weight`.
(2) Opinions with lower `final_weight` should have minimal influence.
(3) When conflicting opinions exist, weigh them according to `final_weight`.
(4) Do not quote reviewer sentences verbatim.

Writing Guidelines:
(1) Write one to three coherent paragraphs.
(2) Maintain a professional academic tone.
(3) Summarize major strengths and weaknesses.
(4) Discuss reviewer disagreements when necessary.
(5) Provide a final recommendation from: accept / weak accept / borderline / weak reject / reject.

Constraints:
(1) Do not mention reviewers explicitly.
(2) Do not output lists or bullet points.
(3) Do not introduce information not supported by the evidence.

Output Format:
Output only the meta-review text without additional explanations or formatting.

Structured Evidence:
`{structured_evidence}`

Table 18: Prompt template for reliability-aware meta-review generation.