MAKING COMPREHENSION EXPLICIT: STEP-WISE INQUIRY FOR AUTOMATED PAPER REVIEWING

Anonymous authors
Paper under double-blind review

ABSTRACT

Peer review requires more than simply producing a critical review; reviewers must question the assumptions of the manuscript, trace its logic, and build a coherent understanding of its claims. However, while existing automated review methods broadly emulate the workflow of human peer review, most do not explicitly model how reviewers progressively build such an understanding before forming their critique. Although this process often manifests in human review as reading annotations, reflection, and fragmented notes, its informal and piecemeal nature does not justify omitting it from automated reviewing. To operationalize this idea, we propose SQUARE, an LLM-based framework that constructs an evidence-grounded understanding of the manuscript through three inquiry phases: sectional inquiry, cross-section inquiry, and comparative novelty assessment. Across rule-based, LLM-as-a-judge, human-grounded review, and human evaluations, SQUARE demonstrates stronger overall performance than capable general-purpose LLMs, specialized review models, and advanced review-agent frameworks. Further analysis suggests that its benefits stem not simply from additional iterative reasoning, but from explicitly organizing and consolidating manuscript evidence into a dedicated representation of understanding. These findings suggest that explicit manuscript understanding, rather than iterative reasoning alone, can serve as a useful intermediate representation for automated paper reviewing.

1 INTRODUCTION

Peer review is a cornerstone of scientific research, serving to assess technical soundness, evaluate conceptual novelty, and safeguard the integrity of the scientific record (Tennant & Ross-Hellauer, 2020). However, the rapid growth in submissions has placed substantial pressure on the peer-review process, as major AI conferences now handle tens of thousands of manuscripts each year (Gropp et al., 2017; Kim et al., 2025). This growing volume can undermine review consistency and lead to overlooked details, driving interest in using Large Language Models (LLMs) to support scientific manuscript evaluation. (Lin et al., 2023; Zhuang et al., 2025).

Existing LLM-based approaches can be broadly categorized into two paradigms: (1) In-context prompting, which uses pre-defined instruction templates to guide LLMs in generating review comments, and extends to multi-agent systems that emulate human expert reasoning or decompose complex review tasks into manageable sub-tasks (Santu et al., 2024; D'Arcy et al., 2024; Lu et al., 2024; 2025; Chang et al., 2025); and (2) Domain-specific fine-tuning, which continually trains open-source models on historical peer-review datasets to enhance their reviewing expertise and domain-specific reasoning (Tan et al., 2024; Yu et al., 2024; Zhu et al., 2025; Zeng et al., 2025).

Despite these advances, automated peer review still struggles to provide sufficiently substantive feedback, often producing generic comments that lack analytical depth and fail to identify critical flaws (Shin et al., 2025). This observation invites a closer look at an often-overlooked foundation of the review process: while most existing approaches focus on high-level simulations of human reviewing, they largely leave manuscript understanding implicit within the LLM's internal reasoning rather than explicitly modeling it (Steer & Ernst, 2021). This leads us to ask: can making manuscript comprehension explicit help LLMs generate more substantive and insightful reviews?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To investigate this idea, we propose SQuARe (**S**tructured **QU**estioning for **A**gentic **RE**viewing), an LLM-based agent framework that explicitly models manuscript comprehension, inspired by the structured inquiry through which human reviewers develop an understanding of a manuscript. Specifically, SQuARe organizes this process into three complementary inquiry phases: (1) *Sectional Inquiry*: The framework conducts sequential query–response cycles at the section level, maintaining structured inquiry notes to capture fine-grained information about the manuscript. (2) *Cross-section Inquiry*: Given the complete manuscript and accumulated inquiry notes, the framework conducts targeted Q&A to uncover overlooked cross-sectional relationships and missing logical dependencies. (3) *Comparative Novelty Assessment*: The framework retrieves related papers and conducts comparative Q&A between the manuscript and existing literature, grounding novelty assessment in cross-referenced evidence. Finally, SQuARe synthesizes the accumulated inquiry notes into a structured review report.

We conduct extensive experiments across four complementary evaluation settings: rule-based, LLM-as-a-judge, human-grounded review, and human evaluation. Compared with capable general-purpose LLMs, specialized review models, and advanced review-agent frameworks, SQuARe demonstrates a consistent overall advantage. Further studies show robust cross-domain generalization and suggest that its gains arise not simply from iterative reasoning, but from explicitly constructing a structured representation of manuscript understanding.

2 Related Work

2.1 Automated Paper Review

Early studies on automated paper reviewing primarily focused on specific subtasks, such as paper acceptance prediction (Kang et al., 2018; Ghosal et al., 2019) and rating recommendation (Kang et al., 2018; Miyao, 2019; Deng et al., 2020), while others explored review generation through knowledge graphs and templates (Wang et al., 2020). With the widespread adoption of LLMs (Achiam et al., 2023; Yang et al., 2025; Guo et al., 2025), recent approaches can be broadly categorized into *domain-specific fine-tuning* and *in-context prompting*.

Domain-specific fine-tuning methods adapt pretrained models to reviewing through additional training. For example, Gao et al. (2024) proposed a two-model framework for manuscript analysis and review generation, while Weng et al. (2025), Zhu et al. (2025), and Zeng et al. (2025) explored hierarchical review simulation, multi-stage reasoning, and reinforcement learning, respectively. However, these approaches require additional training when the underlying base model is updated, incurring substantial costs.

In-context prompting methods perform reviewing through predefined instructions and agentic workflows without modifying model parameters. Representative approaches employ self-reflection (Lu et al., 2024), multi-agent collaboration (Jin et al., 2024; Lu et al., 2025), literature retrieval (Goyal et al., 2026), and recursive question decomposition (Chang et al., 2025), which breaks the overall review task into smaller subtasks and aggregates their results. Despite these diverse strategies, existing approaches primarily focus on review generation, reflection, or task decomposition, leaving manuscript comprehension implicit rather than explicitly modeling it as a dedicated component.

2.2 Progressive and Incremental Reasoning

Recent studies have shown that large language models can reason more effectively when complex problems are decomposed into multiple intermediate steps rather than solved through single-pass generation (Nye et al., 2022; Wei et al., 2022; Kojima et al., 2022; Akyürek et al., 2023; Chen et al., 2025). This has motivated distributed, multi-stage, and iterative reasoning paradigms, in which models progressively update their context, verify intermediate conclusions, and refine reasoning trajectories over multiple steps (Yao et al., 2022; Wang et al., 2023; Madaan et al., 2023; Weng et al., 2023; Yao et al., 2023; Shinn et al., 2024; Pan et al., 2026).

Building on this line of research, *progressive disclosure* (Xu & Yan, 2026) introduces contextual information incrementally, helping reduce context interference in long-context reasoning. SQuARe applies this principle to manuscript comprehension by progressively interrogating different sections, accumulating structured inquiry traces, and integrating information across sections for subsequent reasoning and review synthesis.

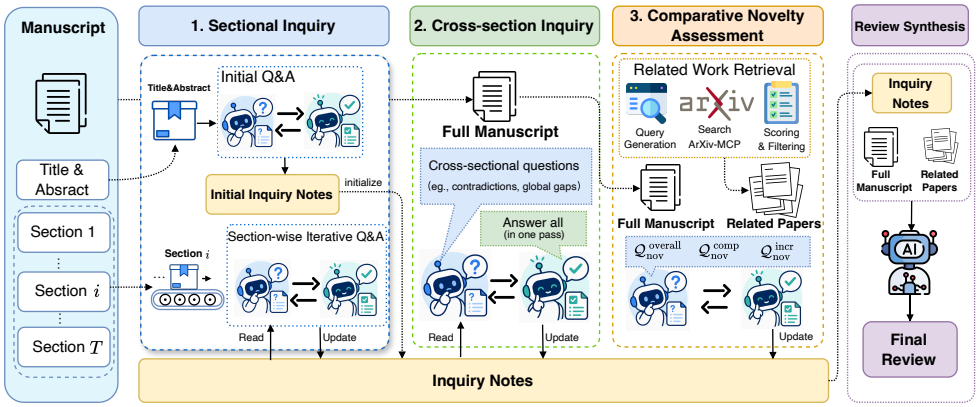

Figure 1: Overview of SQUARE. The framework performs step-wise inquiry over the manuscript through section-level, cross-section, and novelty-focused question answering, accumulates evidence in the structured inquiry notes, and synthesizes the final review from the aggregated evidence.

3 METHOD

As shown in Figure 1, our proposed SQUARE comprises three sequential inquiry stages: *Sectional Inquiry*, *Cross-section Inquiry*, and *Comparative Novelty Assessment*. Each stage employs two collaborative agents—a *Question Generation Agent* $\mathcal{A}_q$ and an *Answer Generation Agent* $\mathcal{A}_a$—where $\mathcal{A}_q$ formulates context-aware questions conditioned on the manuscript content and accumulated context, and $\mathcal{A}_a$ produces the corresponding answers. The resulting question–answer pairs are cumulatively recorded in the *Inquiry Notes* $\mathcal{I}$, which serves as an evolving, structured memory throughout the review process. Once all three stages are complete, the framework synthesizes the final review from the aggregated inquiry notes.

3.1 SECTIONAL INQUIRY

This stage captures fine-grained, section-level information by performing question answering over each section of the manuscript sequentially. Let $\mathcal{M}$ denote the manuscript under review, modeled as an ordered sequence of sections:

$$\mathcal{M} = (s_0, s_1, \ldots, s_T), \tag{1}$$

where s_0 comprises the title and abstract, each s_i ($i \geq 1$) denotes the i-th body section, and T is the total number of sections.

Initial Q&A. We first construct initial inquiry notes from s_0. The question generation agent produces an initial set of questions aimed at establishing a high-level understanding of the manuscript:

$$\mathcal{Q}_0 = \mathcal{A}_q(s_0) = \{q_j\}_{j=1}^{M}, \tag{2}$$

where $M \leq M_{\max}$ and $M_{\max}$ is a predefined upper bound on the number of initial questions. The answer generation agent then produces the corresponding answers jointly, conditioned on s_0 and $\mathcal{Q}_0$:

$$\mathcal{R}_0 = \mathcal{A}_a(s_0, \mathcal{Q}_0) = \{a_j\}_{j=1}^{M}. \tag{3}$$

The initial inquiry notes are then defined as $\mathcal{I}_0 = \{(q_j, a_j)\}_{j=1}^{M}$.

Section-wise Iterative Q&A. Building on the initial notes, the framework iterates over each subsequent section to progressively deepen understanding and surface potential issues. For each section s_i ($i \geq 1$), the question generation agent formulates a new set of questions conditioned on the current section and the accumulated inquiry notes:

$$\mathcal{Q}_i = \mathcal{A}_q(s_i, \mathcal{I}_{i-1}) = \{q_k\}_{k=1}^{N_i}, \tag{4}$$

where $N_i \leq N_{\max}$ and $N_{\max}$ is the predefined maximum number of questions per section. The generated questions fall into two categories: (1) *independent questions* that arise directly from the content of s_i and (2) *derived questions* that refine or extend previous entries in $\mathcal{I}_{i-1}$ in light of the newly encountered material. To preserve a clear logical hierarchy, we adopt a *Tractatus*-style decimal numbering scheme. Independent questions are assigned fresh top-level indices (e.g. `[3]`), whereas derived questions are numbered as sub-entries of their parent (e.g. `[1.2]` for the second refinement of question `[1]`, or `[1.2.1]` for a further sub-question under `[1.2]`). This hierarchical indexing makes the logical provenance of each question explicit, ensuring traceability throughout the inquiry process.

The newly generated questions are appended to the inquiry notes with their answers initialized to a null placeholder $\varnothing$, forming the augmented notes:

$$\widetilde{\mathcal{I}}_i = \mathcal{I}_{i-1} \cup \{(q_k, \varnothing)\}_{k=1}^{N_i}. \tag{5}$$

The answer generation agent then receives the current section s_i along with the full augmented notes $\widetilde{\mathcal{I}}_i$ and produces answers for all entries in a single pass:

$$\mathcal{R}_i = \mathcal{A}_a(s_i, \widetilde{\mathcal{I}}_i) = \{a'_j\}_{j=1}^{|\widetilde{\mathcal{I}}_i|}, \tag{6}$$

where each a'_j corresponds to the j-th entry in $\widetilde{\mathcal{I}}_i$: for previously answered entries ($a_j \neq \varnothing$), the agent retains the existing answer ($a'_j = a_j$); for unanswered entries ($a_j = \varnothing$), it either supplies a substantive answer grounded in s_i or preserves $\varnothing$ when the section does not provide sufficient evidence. The updated inquiry notes are then constructed by pairing each question with its corresponding answer:

$$\mathcal{I}_i = \left\{(q_j,\, a'_j)\right\}_{j=1}^{|\widetilde{\mathcal{I}}_i|}. \tag{7}$$

This mechanism enables questions raised in earlier sections to be progressively resolved as additional content becomes available, mirroring how a human reviewer revisits open questions while reading subsequent parts of a manuscript.

After processing all T sections, the Sectional Inquiry stage yields the cumulative inquiry notes $\mathcal{I}_T = \{(q_j, a'_j)\}_{j=1}^{K}$, where $K = M + \sum_{i=1}^{T} N_i$ is the total number of accumulated questions. The inquiry notes record every question raised throughout the manuscript along with its resolved or still-pending answer, serving as input to the subsequent Cross-section Inquiry stage.

3.2 Cross-section Inquiry

The Sectional Inquiry stage processes each section independently, which may leave cross-section issues undetected—for instance, logical inconsistencies between sections or globally unsupported claims. This stage complements the previous one by shifting the scope from individual sections to the entire manuscript $\mathcal{M}$.

Following the same question-then-answer protocol, $\mathcal{A}_q$ receives $\mathcal{M}$ and the accumulated notes $\mathcal{I}_T$, and generates questions targeting aspects not yet adequately covered—particularly cross-section concerns such as contradictions between claims and experimental evidence, missing baseline comparisons, and methodological gaps that span multiple sections:

$$\mathcal{Q}_{\text{cross}} = \mathcal{A}_q(\mathcal{M}, \mathcal{I}_T) = \{q_k\}_{k=1}^{P}, \tag{8}$$

where $P \leq P_{\max}$, and the *Tractatus*-style numbering is maintained. These questions are appended to $\mathcal{I}_T$ with null placeholders, forming the augmented notes $\widetilde{\mathcal{I}}_{\text{cross}} = \mathcal{I}_T \cup \{(q_k, \varnothing)\}_{k=1}^{P}$.

$\mathcal{A}_a$ then takes $\mathcal{M}$ and $\widetilde{\mathcal{I}}_{\text{cross}}$ as input and addresses all unanswered entries in a single pass:

$$\mathcal{R}_{\text{cross}} = \mathcal{A}_a(\mathcal{M}, \widetilde{\mathcal{I}}_{\text{cross}}) = \{a'_j\}_{j=1}^{|\widetilde{\mathcal{I}}_{\text{cross}}|}. \tag{9}$$

With access to the full manuscript, questions that were previously unanswerable from individual sections can now be resolved via cross-referencing. The agent produces a complete answer for each question that $\mathcal{M}$ sufficiently addresses; for those that remain irresolvable, it provides a partial response summarizing the available evidence and explaining why the question cannot be fully answered. The resulting inquiry notes are:

$$\mathcal{I}_{\text{cross}} = \left\{(q_j,\, a'_j)\right\}_{j=1}^{|\widetilde{\mathcal{I}}_{\text{cross}}|}, \tag{10}$$

where every entry now carries either a complete answer or an explicit account of insufficiency, ensuring comprehensive coverage before the final review generation.

3.3 Comparative Novelty Assessment

A thorough assessment of a manuscript requires situating it within the broader research landscape. This section further complements the evaluation by retrieving and examining related work, providing a more complete perspective on the paper's novelty.

Following Zeng et al. (2025), we adopt the ArXiv-MCP[1] server as the retrieval backend for related work discovery. An LLM first generates a small set of diverse search queries from the manuscript's title and abstract, each covering a distinct facet of the research topic. The queries are issued to the server and the union of returned results forms a raw candidate pool. The LLM then scores each candidate for relevance to the manuscript on a $[0, 100]$ scale; after de-duplication and thresholding, the top-ranked papers are retained to form the related paper set $\mathcal{P} = \{p_k\}_{k=1}^{L}$, where $L \leq L_{\max}$.

Given $\mathcal{M}$ and $\mathcal{P}$, $\mathcal{A}_q$ generates novelty-oriented questions organized into three complementary categories:

$$\mathcal{Q}_{\text{nov}} = \mathcal{A}_q(\mathcal{M}, \mathcal{P}) = \mathcal{Q}_{\text{nov}}^{\text{overall}} \cup \mathcal{Q}_{\text{nov}}^{\text{comp}} \cup \mathcal{Q}_{\text{nov}}^{\text{incr}}, \tag{11}$$

where $\mathcal{Q}_{\text{nov}}^{\text{overall}}$ probes the high-level independence and significance of the contribution, $\mathcal{Q}_{\text{nov}}^{\text{comp}}$ references specific related papers to examine methodological or conceptual overlap, and $\mathcal{Q}_{\text{nov}}^{\text{incr}}$ scrutinizes whether the claimed contributions are incremental variations or re-packaging of prior work. $\mathcal{A}_a$ then takes the full content of $\mathcal{M}$, the abstracts of $\mathcal{P}$, and $\mathcal{Q}_{\text{nov}}$ as input, producing the corresponding answers in a single pass.

$$\mathcal{R}_{\text{nov}} = \mathcal{A}_a(\mathcal{M}, \mathcal{P}, \mathcal{Q}_{\text{nov}}) = \{a_j\}_{j=1}^{|\mathcal{Q}_{\text{nov}}|}. \tag{12}$$

The resulting novelty assessment notes $\mathcal{I}_{\text{nov}} = \{(q_j, a_j)\}_{j=1}^{|\mathcal{Q}_{\text{nov}}|}$ pairs each question with its answer.

After the three inquiry stages, the accumulated notes are merged into unified inquiry notes $\mathcal{I} = \mathcal{I}_{\text{cross}} \cup \mathcal{I}_{\text{nov}}$. A review synthesis module then produces the final structured review:

$$\mathcal{V} = \text{Synthesize}(\mathcal{M}, \mathcal{I}, \mathcal{P}), \tag{13}$$

where $\mathcal{V}$ covers summary, strengths, weaknesses, questions, and numerical ratings.

4 Experiments

4.1 Experimental Settings

Implementation Details. We use the Gemini-2.5-Flash API as the backbone LLM for all agents in SQuARe, with the temperature set to 0 and the maximum output length set to 32,768 tokens. For the framework hyperparameters, we set $M_{\max} = N_{\max} = P_{\max} = L_{\max} = 5$ and use a relevance threshold of 70 for related work filtering. In practice, when no related paper survives the filtering stage (i.e., $\mathcal{P} = \emptyset$), we skip the Comparative Novelty Assessment and directly synthesize the final review from $\mathcal{I}_{\text{cross}}$ and $\mathcal{M}$.

Dataset. We construct our evaluation dataset by randomly sampling 200 papers from the DeepReview-Test[2] (Zhu et al., 2025), more than double the 80 papers used in the recent TreeReview study (Chang et al., 2025). Detailed information about the dataset is provided in Appendix A.

Baselines. Following the two paradigms identified in our related work, we compare against baselines from domain-specific fine-tuning and in-context prompting, and select representative methods from each. **(1) Domain-specific fine-tuning.** Within this category, we consider two state-of-the-art publicly available models, CycleReviewer-8B (Weng et al., 2025) and DeepReviewer-7B (Zhu et al., 2025). Note that both are trained to regress the *average* human rating and thus emit continuous scores, giving them a structural advantage on score-based metrics over methods restricted to integer predictions. **(2) In-context prompting.** We further distinguish two sub-types: **(a) Direct LLM review**, which prompts a single model to produce the complete review in one pass. We cover seven

[1] https://github.com/blazickjp/arxiv-mcp-server

[2] Deepreview-Test is released under the CycleResearcher License, which permits non-commercial research use.

Table 1: Rule-based evaluation results. Arrows indicate metric direction (↓: lower is better; ↑: higher is better). **Bold** marks the best score within each category.

Model	Rating		Soundness		Presentation		Contribution		Rating Spearman↑
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	
Domain-specific fine-tuning									
CycleReviewer	1.07	1.75	**0.35**	**0.21**	0.42	**0.28**	0.39	0.24	0.29
DeepReviewer	**1.02**	**1.64**	0.38	0.24	**0.41**	**0.28**	**0.38**	**0.23**	**0.43**
In-context prompting — direct LLM review									
GPT-5	2.49	7.47	0.95	1.19	0.55	0.47	1.39	1.23	0.39
GPT-4.1	2.34	6.80	0.90	1.05	0.48	0.37	1.24	1.83	0.33
Claude Sonnet 4	1.05	1.69	0.43	0.31	0.48	0.36	0.57	0.48	0.39
Gemini 2.5 Flash	2.23	6.31	0.91	1.09	0.94	1.18	2.23	6.31	0.35
DeepSeek-R1	1.98	5.27	0.91	1.17	0.66	0.68	1.30	2.03	0.25
DeepSeek-V3.1	1.83	4.40	1.10	1.45	0.86	1.05	1.50	2.44	**0.45**
Qwen3-32B	1.92	5.00	1.18	1.66	0.73	0.86	1.23	1.89	0.20
In-context prompting — multi-agent review									
AI-Scientist	1.95	5.37	0.75	0.88	0.66	0.68	0.72	0.82	0.23
TreeReview	1.37	2.96	0.52	0.45	0.60	0.54	0.64	0.60	0.03
SQuARe (Ours)	**1.03**	**1.67**	**0.42**	**0.28**	**0.47**	**0.34**	**0.49**	**0.36**	0.41

models spanning three families: closed-source frontier models—GPT-5 (`gpt-5-chat`), GPT-4.1 (`gpt-4.1-2025-04-14`), Claude Sonnet 4 (`claude-sonnet-4-20250514`), and Gemini 2.5 Flash (`gemini-2.5-flash`); a reasoning-specialized model—DeepSeek-R1 (`deepseek-r1-250528`); and open-weight models—DeepSeek-V3.1 (`deepseek-v3.1`) and Qwen3 (`qwen3-32b`). **(b) Multi-agent review**, which orchestrates multi-step workflows or multi-agent collaboration to emulate reviewing. We include AI-Scientist (Lu et al., 2024) and TreeReview (Chang et al., 2025). For a fair comparison, both multi-agent baselines are powered by the same Gemini 2.5 Flash backbone as SQuARe.

CycleReviewer and DeepReviewer are run on a single NVIDIA A100 GPU (40,GB). AI-Scientist and TreeReview follow their official implementations and default configurations, with TreeReview's auxiliary LLaMA-3.1-8B model hosted on a single NVIDIA RTX 4090 GPU (24,GB). Failed or non-terminating runs are repeated until a complete review is obtained.

4.2 Results

Rule-based Results. We evaluate predicted ratings using MAE, MSE, and Spearman's rank correlation coefficient (ρ); results are shown in Table 1. SQuARe achieves near-best performance among in-context prompting methods, substantially outperforming its backbone, Gemini 2.5 Flash (MAE/MSE: $1.03/1.67$ vs. $2.23/6.31$), and approaching the performance of Claude Sonnet 4, a higher-cost general-purpose model ($1.05/1.69$). Both outperform the leave-one-reviewer-out human-agreement reference (MAE 1.20), indicating closer agreement with reviewer consensus. On the same backbone, SQuARe reduces Rating MAE by 25% over TreeReview and 47% over AI-Scientist, while matching fine-tuned specialists on Rating (MAE/MSE: $1.03/1.67$ vs. $1.02/1.64$ and $1.07/1.75$) and outperforming CycleReviewer-8B in MAE, despite the latter benefiting from direct regression of continuous reviewer-average scores. For a more comparable evaluation, Appendix B reports results after rounding the continuous predictions of fine-tuned methods to the nearest integer on the rating scale used by SQuARe, under which SQuARe demonstrates a clearer advantage.

LLM-as-a-judge Result. Beyond rule-based rating metrics, we further assess the textual quality of the generated reviews through an LLM-as-a-judge (Liu et al., 2023; Zheng et al., 2023; Gu et al., 2024) protocol. Inspired by the evaluation setup of Zhu et al. (2025), we perform pairwise comparisons between reviews produced by SQuARe and each baseline along three fine-grained dimensions—*Constructive Value*, *Analytical Depth*, and *Technical Accuracy*—together with an *Overall Judgment* dimension that gives a holistic preference (see Appendix C.1 for details). All pairwise judgments are made by Gemini-2.5-Pro, a stronger model than the Gemini-2.5-Flash backbone shared by SQuARe. The results are shown in Table 2. SQuARe obtains an overall win rate above 79.5% against every baseline, covering both domain-specific fine-tuned reviewers and in-context prompting methods,

Table 2: LLM-as-a-judge pairwise evaluation (%). Win/Lose rates of SQuARE against each baseline.

Baseline (SQuARE *vs.*)	Overall Judgment		Constructive Value		Analytical Depth		Technical Accuracy	
	Win↑	Lose↓	Win↑	Lose↓	Win↑	Lose↓	Win↑	Lose↓
Domain-specific fine-tuning								
CycleReviewer	**100.0**	0.0	**100.0**	0.0	**100.0**	0.0	**100.0**	0.0
DeepReviewer	**100.0**	0.0	**99.0**	0.0	**99.5**	0.0	**99.5**	0.0
In-context prompting — direct LLM review								
GPT-5	**98.0**	2.0	**95.5**	2.0	**97.5**	2.5	**89.5**	1.5
GPT-4.1	**97.5**	2.5	**95.0**	3.0	**98.0**	2.0	**88.5**	2.5
Claude Sonnet 4	**100.0**	0.0	**97.5**	1.0	**100.0**	0.0	**93.5**	1.5
Gemini 2.5 Flash	**89.5**	9.0	**75.0**	17.0	**89.5**	8.5	**71.0**	10.0
DeepSeek-R1	**93.0**	7.0	**87.0**	9.0	**94.0**	5.5	**80.5**	7.0
DeepSeek-V3.1	**92.5**	6.5	**87.5**	8.5	**93.0**	7.0	**77.5**	6.0
Qwen3-32B	**100.0**	0.0	**97.5**	0.0	**100.0**	0.0	**97.0**	0.5
In-context prompting — multi-agent review								
AI-Scientist	**79.5**	20.0	**77.5**	15.0	**78.5**	21.0	**62.5**	20.5
TreeReview	**96.0**	3.5	**88.0**	6.5	**96.5**	3.5	**90.0**	2.0

Table 3: Ground-truth evaluation (%). Baselines are sorted by SN-F1.

System	SN-P ↑	SN-R ↑	SN-F1 ↑
Qwen3-32B	43.19	43.36	43.27
Gemini 2.5 Flash	41.94	44.12	43.00
TreeReview	43.38	41.58	42.46
Claude Sonnet 4	41.05	43.82	42.39
DeepReview-7B	42.22	42.02	42.12
DeepSeek-V3.1	**43.54**	39.02	41.15
GPT-4.1	38.39	43.05	40.59
AI-Scientist	39.33	41.28	40.28
CycleReviewer-8B	42.18	37.97	39.96
GPT-5	37.26	41.03	39.05
SQuARE (Ours)	41.49	**47.07**	**44.11**

and the preference holds across all four dimensions, with the narrowest margins on Technical. We note, however, that the uniformly high win rates suggest that LLM judges may partly favor surface properties such as length and formatting; we therefore treat this result as indicative rather than conclusive, and corroborate it with human evaluation and format-invariant ground-truth evaluation. We further analyze potential judge-model bias and review length in Appendix D. Specifically, we repeat the evaluation using two cross-family judges, DeepSeek-V3.2 and GPT-5.4, and examine the relationship between review length and win rate. The results suggest that the observed preference cannot be explained by judge-model choice or verbosity alone.

Human-grounded Review Evaluations. We further evaluate how well the weaknesses identified by each system align with those raised by human reviewers, following Chang et al. (2025). Specifically, we decompose both human reviews and system outputs into atomic weakness statements, construct a semantic similarity matrix using `all-mpnet-base-v2`, and compute SN-Precision, SN-Recall, and SN-F1 (Lou et al., 2024). As shown in Table 3, SQuARE achieves the highest SN-F1 (44.11), driven primarily by its leading recall (47.07), indicating broader coverage of weaknesses identified by human reviewers, while maintaining comparable precision. Because the metric directly matches atomic weaknesses against human-written references, it is insensitive to formatting and stylistic variation, providing a complementary measure to the LLM-as-a-judge evaluation.

Human Evaluation Results. We conduct a limited-scale pairwise human preference evaluation among SQuARE, Gemini-2.5-Flash, Claude-Sonnet-4, AI-Scientist, and DeepReview-7B. Six volunteers with prior peer-review experience evaluate 30 randomly sampled test manuscripts (details in Appendix C.2). As shown in Figure 2, SQuARE receives a higher preference rate than all four baselines, with the largest preference margins observed against AI-Scientist and DeepReview-7B.

Additional analyses are provided in the appendix: controlling for external literature retrieval in Appendix E, evaluating robustness across backbone models in Appendix F, and reporting statistical significance analyses and case studies in Appendices G and H, respectively.

4.3 Ablation Study

To quantify the contribution of each inquiry stage, we remove one stage at a time from SQuARE while keeping all other components fixed, and evaluate each variant on the same 200 papers. We report Rating MAE and MSE. As shown in Figure 3, removing any stage degrades both metrics relative to the full system, suggesting that *Sectional Inquiry*, *Cross-section Inquiry*, and *Comparative Novelty Assessment* provide complementary evidence and jointly contribute to review quality.

Among the three stages, *Sectional Inquiry* has the largest impact: removing it produces the largest increase in MAE and a substantial rise in MSE, highlighting the importance of fine-grained, section-level grounding. Removing *Comparative Novelty Assessment* leads to a moderate increase in MAE but the highest MSE among all variants, suggesting that comparative evidence from related work helps reduce larger rating errors. Finally, removing *Cross-section Inquiry* results in smaller but

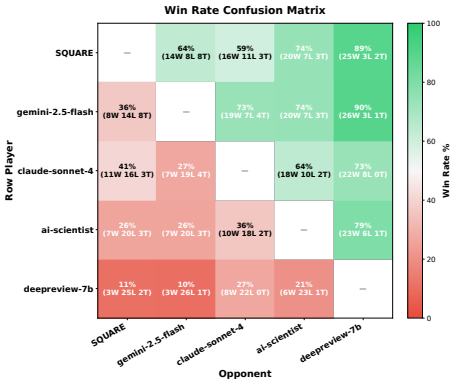

Figure 2: Human pairwise win-rate matrix.

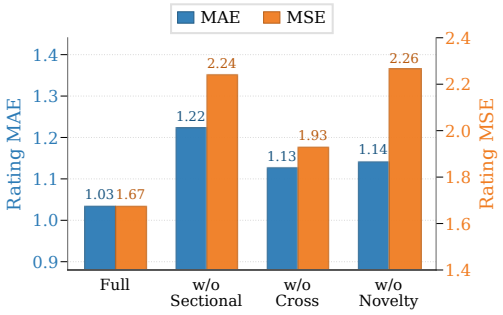

Figure 3: Ablation of the three inquiry stages.

consistent degradations on both metrics, indicating that probing inter-section relationships and missing dependencies provides additional evidence beyond section-level grounding.

4.4 Analysis

To examine the role of explicitly modeling manuscript comprehension, we construct two controlled baselines. *Interleaved Synthesis* retains the three-stage inquiry process but incorporates each round's findings directly into an evolving review rather than maintaining separate inquiry notes, while using a token budget comparable to SQuARE. *Interleaved Direct Revision* further removes the questioning process and directly revises the review as new manuscript content is revealed. We compare these variants with SQuARE using overall Rating MAE/MSE and Overall Judgment under LLM-as-a-judge evaluation.

As shown in Table 4, both variants underperform SQuARE, with the QA-free variant showing a substantially larger degradation. The relatively large gap from *Interleaved Synthesis* is particularly revealing: despite following the same three-stage schedule and using a comparable token budget, directly folding inquiry findings into an evolving review is markedly less effective than maintaining them as separate inquiry notes. This suggests that the gain from inquiry is fundamentally representational rather than merely computational—the crucial factor is not how many rounds the model reasons, but what representation is carried across those rounds. A review draft is inherently judgment-laden, so using it as the intermediate state entangles evidence gathering with ongoing evaluation; in contrast, the inquiry notes preserve a dedicated, evidence-oriented representation of the manuscript that can be progressively refined and integrated. The further collapse of *Interleaved Direct Revision* shows that, without this intermediate extraction and consolidation process, repeated revision alone provides little benefit. Taken together, the results suggest that effective automated reviewing may require not just iterative reasoning, but an explicit intermediate representation in which manuscript evidence is organized before it is transformed into evaluative judgments.

Table 4: Comparison of different review synthesis strategies.

	Rating		Preference vs. SQuARE	
Method	MAE ↓	MSE ↓	SQuARE Win ↑	SQuARE Lose ↓
SQuARE	**1.03**	**1.67**	–	–
Interleaved Synthesis	1.51	3.51	67.5%	31.5%
Interleaved Direct Revision	2.86	9.88	91.5%	8.5%

4.5 Domain Robustness Study

To assess the domain robustness of SQuARE, we curate 40 held-out *Nature Communications* papers from climate science (13), energy (14), and engineering (13). Using LLM-as-a-judge evaluation, we compare SQuARE with four strong baselines from Section 4.1: Gemini-2.5-Flash, DeepSeek-R1,

Table 5: Per-paper average token usage of SQUARE and representative multi-agent review frameworks.

Method	Input	Output	Total
MARG	2,192,910	121,141	2,314,052
MARG-Base	963,027	44,581	1,007,608
TreeReview	419,929	39,039	458,968
AI-Scientist	191,792	20,937	212,729
SQUARE	208,520	26,289	234,809

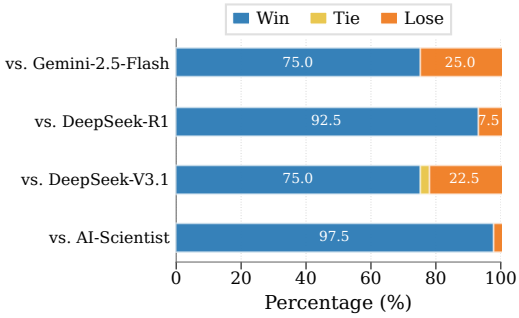

Figure 4: LLM-as-a-judge pairwise evaluation on 40 held-out *Nature Communications* papers.

DeepSeek-V3.1, and AI-Scientist. To reduce domain-specific biases, we disable related-work retrieval in SQUARE and remove the original ML-paper few-shot examples from AI-Scientist. The Overall Judgment results are shown in Figure 4. SQUARE maintains win rates of at least 75% across these out-of-domain papers, indicating robust cross-domain performance beyond the settings used for development.

4.6 Computational Cost

In this section, we analyze the computational cost of SQUARE by comparing its average input, output, and total token consumption per review with MARG (D'Arcy et al., 2024), MARG-Base, TreeReview (Chang et al., 2025), and AI-Scientist (Lu et al., 2024). Token statistics for some baselines are taken from Chang et al. (2025). As shown in Table 5, SQUARE uses an order of magnitude fewer tokens than MARG and MARG-Base, fewer than TreeReview, and a comparable number to AI-Scientist, indicating strong token efficiency.

5 Conclusion

In this work, we investigate an often-overlooked component of automated peer review: the process by which a reviewer builds an understanding of a manuscript before forming a critique. While most existing methods treat this process as implicit internal reasoning, we argue that it is useful to model it explicitly. Based on this view, we introduce SQUARE, an LLM-based review framework that constructs an evidence-grounded representation of manuscript understanding through sectional inquiry, cross-section inquiry, and comparative novelty assessment. Our analyses further suggest that the observed gains arise not simply from additional iterative reasoning, but from explicitly organizing and consolidating manuscript evidence into a structured representation of understanding. More broadly, explicit manuscript understanding may serve as a useful intermediate representation that complements existing workflow-oriented review agents, offering an extensible building block for future automated review frameworks.

AI use statement

In this work, we used generative AI tools solely to assist with writing, including language polishing and improving the clarity and readability of the manuscript. We did not use generative AI tools for research ideation, methodology design, implementation, data analysis, result interpretation, or the generation of scientific claims. All AI-assisted text was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

SQUARE is intended to assist, rather than replace, human peer review. Its outputs should be treated as supplementary material and should not serve as the sole basis for acceptance, rejection, or scoring decisions. We also acknowledge that LLM-based review systems may inherit biases

from their training data and may produce uneven assessments across research topics, methodologies, communities, or writing styles. Accordingly, we emphasize the need for human oversight, careful verification of model outputs, and continued evaluation of potential biases and failure modes when deploying automated review systems.

Reproducibility statement

To support reproducibility, we provide the complete prompts used in all stages of SQuARe in the appendix. All experiments are conducted on publicly available datasets, with the corresponding experimental settings and evaluation details described in the main paper and appendix.

References

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

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15662–15693, 2025.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

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

Zhongfen Deng, Hao Peng, Congying Xia, Jianxin Li, Lifang He, and Philip S Yu. Hierarchical bi-directional self-attention networks for paper review rating recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6302–6314, 2020.

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

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1120–1130, 2019.

Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. Scholarpeer: A context-aware multi-agent framework for automated peer review. *arXiv preprint arXiv:2601.22638*, 2026.

Robert E Gropp, Scott Glisson, Stephen Gallo, and Lisa Thompson. Peer review: A system under stress. *BioScience*, 67(5):407–410, 2017.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1208–1226, 2024.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long papers)*, pp. 1647–1661, 2018.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. In *International Conference on Machine Learning*, pp. 81634–81651. PMLR, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Information fusion*, 98:101830, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, et al. Aaar-1.0: Assessing ai's potential to assist research. *arXiv preprint arXiv:2410.22394*, 2024.

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

Kai Lu, Shixiong Xu, Jinqiu Li, Kun Ding, and Gaofeng Meng. Agent reviewers: Domain-specific multimodal agents with shared memory for paper review. In *International Conference on Machine Learning*, pp. 40803–40830. PMLR, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yusuke Miyao. Does my rebuttal matter? insights from a major nlp conference. In *Proceedings of NAACL-HLT*, pp. 1274–1290, 2019.

Maxwell Nye, Anders Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, Quoc V. Le, and Denny Zhou. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2022.

Liangming Pan, Jason Liang, Jiaran Ye, Minglai Yang, Xinyuan Lu, and Fengbin Zhu. Opening the black box: A survey on the mechanisms of multi-step reasoning in large language models. *arXiv preprint arXiv:2601.14270*, 2026.

Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, et al. Prompting llms to compose meta-review drafts from peer-review narratives of scholarly manuscripts. *arXiv preprint arXiv:2402.15589*, 2024.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Automatically evaluating the paper reviewing capability of large language models. *arXiv e-prints*, pp. arXiv–2502, 2025.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. *URL https://arxiv.org/abs/2303.11366*, 8, 2024.

Philip J Steer and Sabine Ernst. Peer review-why, when and how, 2021.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.

Jonathan P Tennant and Tony Ross-Hellauer. The limitations to our understanding of peer review. *Research integrity and peer review*, 5(1):6, 2020.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 384–397, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575. Association for Computational Linguistics, 2023.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Renjun Xu and Yang Yan. Agent skills for large language models: Architecture, acquisition, security, and the path forward. *arXiv preprint arXiv:2602.12430*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184, 2024.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, et al. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16942–16954, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, 2025.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332, 2025.

A Dataset Analysis

We provide a descriptive analysis of the evaluation dataset used in our experiments.

Overview. The test set comprises 200 peer-reviewed submissions with binary editorial decisions: 74 accepted (37%) and 126 rejected (63%). Each submission is accompanied by 3–7 reviewer reports (mean 4.11, median 4).

Rating distribution. Figure 5 shows the per-review rating distribution grouped by final decision. Ratings for rejected papers concentrate at 3 and 5, while accepted papers peak at 6 and 8, confirming a clear separation. The mean rating per paper is 5.26 (median 5.25; interquartile range [4.40, 6.25]).

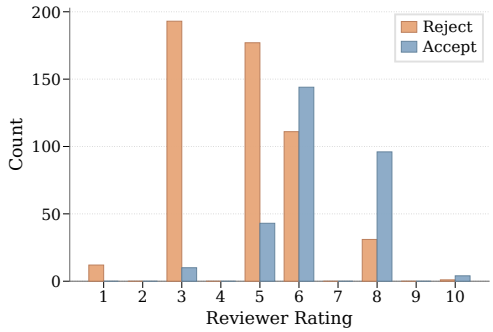

Figure 5: Distribution of individual reviewer ratings in the test subset, grouped by the final editorial decision (*Accept* vs. *Reject*).

Text length statistics. The bodies of the paper average 41,864 characters (median 39,470; 90th percentile 52,585). Individual reviews average 6,506 characters (median 6,249; 90th percentile 10,658), reflecting substantive and detailed feedback.

B Comparison under a Discrete Rating Scale

The fine-tuned baselines (CycleReviewer and DeepReviewer) regress continuous average-reviewer scores, whereas SQuARe predicts integer ratings. This difference can favor continuous-output models on MAE/MSE, as fractional predictions can more closely approximate continuous gold scores than integer predictions. To examine the extent of this advantage, we additionally round the baseline predictions to the nearest integer and re-evaluate them on the same rating scale.

As shown in Table 6, rounding substantially increases the MAE and MSE of both fine-tuned baselines. Under the resulting discrete-rating setting, SQuARe achieves the best Rating MAE/MSE and matches or outperforms the baselines on Soundness. Overall, rounding the continuous predictions of the fine-tuned baselines further highlights the performance advantage of SQuARe on Rating.

C Evaluation Setting Details

C.1 LLM-as-a-judge Evaluation Settings

Our LLM-as-a-judge evaluation protocol largely follows that of Zhu et al. (2025). We employ a judge model to conduct pairwise comparative evaluations between reviews generated by SQuARe and those produced by each baseline system. For every evaluation instance, the judge is presented with (1) the original paper and (2) a pair of reviews from two different systems in randomized order, where

Table 6: MAE / MSE under continuous vs. discretized scoring.

	Rating		Soundness		Presentation		Contribution	
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
Continuous (original)								
DeepReviewer	1.02	1.64	0.38	0.24	0.41	0.28	0.38	0.23
CycleReviewer	1.07	1.75	0.35	0.21	0.42	0.28	0.39	0.24
Discretized (rounded to integer)								
DeepReviewer	1.09	1.81	0.45	0.33	**0.46**	**0.34**	**0.45**	**0.31**
CycleReviewer	1.14	1.98	**0.42**	**0.28**	0.47	**0.34**	0.50	0.36
SQuARe	**1.03**	**1.67**	**0.42**	**0.28**	0.47	**0.34**	0.49	0.36

each review comprises a summary, strengths, weaknesses, and questions. The judge then renders one of three verdicts: *win*, *lose*, or *tie*.

The assessment is conducted along the following three sub-dimensions:

Technical Accuracy. Whether claims are consistent with the paper content, evidence is properly interpreted, technical assessments are valid, and critiques are well-supported.

Constructive Value. Whether the feedback is actionable, suggestions are specific and feasible, criticism is balanced with acknowledgment of strengths, and authors can clearly understand how to improve.

Analytical Depth. Whether key aspects are thoroughly examined, analysis is appropriately detailed, important elements are addressed, and the assessment is comprehensive.

Based on these three sub-dimensions, the judge also provides an *Overall Judgment*.

In practice, we adopt Gemini-2.5-Pro as the judge model, with `reasoning_max_tokens` set to 2048. The full prompt is provided in Figure 6.

C.2 Human Evaluation Settings

We recruit six annotators, all of whom are PhDs or PhD candidates with prior reviewing experience. From the test set, we randomly sample 30 papers and assign 5 papers to each annotator. For every paper, all $\binom{5}{2}=10$ pairs of system-generated reviews are presented in anonymized and randomized order, and the annotator selects the preferred review, or marks the pair as a *tie* when the two reviews are of comparable quality.

Before annotation, each annotator receives written guidelines. Reviews are evaluated along five dimensions: *Accuracy*, *Coverage*, *Specificity*, *Constructiveness*, and *Clarity*. Annotators are instructed to assess the *quality of the review*, rather than the underlying paper, based solely on the manuscript content; to avoid using length, formatting, or stylistic cues as proxies for quality; and to treat reviews as blind, without inferring their human or model origin. When two reviews are difficult to distinguish, annotators may assign a *tie* and briefly explain the conflicting evidence. For each comparison, annotators also provide an *Overall Judgment*, a confidence score on a 1–5 Likert scale, and a free-form justification, which is required for ties.

D Bias Analysis

Judge-model Bias Analysis. To assess the robustness of the preference for SQuARe to the choice of judge model, we replace the Gemini-2.5-Pro judge used in the main experiments with DeepSeek-V3.2, which belongs to a different model family from the Gemini-2.5-Flash backbone of SQuARe. As shown in Table 7, SQuARe remains preferred over all baselines across all evaluation dimensions. Although the win rates against some baselines, such as Gemini 2.5 Flash and DeepSeek-R1, are lower than those obtained with Gemini-2.5-Pro, the overall pattern is consistent: Overall Judgment win

rates range from 70.5% to 100.0%, with lose rates below 30% in all comparisons. This suggests that the observed preference is not specific to the original judge model.

We further evaluate several representative baselines using GPT-5.4 as a second cross-family judge on the same 200 papers. As reported in Table 8, SQUARE again achieves higher preference rates, with Overall Judgment win rates ranging from 68.5% against Gemini 2.5 Flash to 100.0% against DeepReview-7B. Taken together, the results across three judges from distinct model families indicate that the preference for SQUARE is robust to the choice of judge model.

Length Bias Analysis. We also examine whether the evaluation is affected by verbosity bias, whereby longer reviews may be favored irrespective of quality. Table 9 reports the average character count of the review text provided to the judge. SQUARE (8,969 characters on average) is shorter than DeepReviewer (15,181) but attains a 100% win rate against it; conversely, CycleReviewer is substantially shorter (1,752 characters), while SQUARE also achieves a 100% win rate. Across all 11 baselines, the Spearman correlation between the length ratio (SQUARE/baseline) and win rate is $\rho = -0.23$. This weak correlation provides little evidence that review length systematically drives the observed preferences.

Table 7: Evaluation Results Using DeepSeek V3.2 as the Judge Model.

Baseline (SQUARE *vs.*)	**Overall Judgment**		**Constructive Value**		**Analytical Depth**		**Technical Accuracy**	
	Win ↑	Lose ↓	Win ↑	Lose ↓	Win ↑	Lose ↓	Win ↑	Lose ↓
Domain-specific fine-tuning								
CycleReviewer	**100.0**	0.0	**100.0**	0.0	**100.0**	0.0	**100.0**	0.0
DeepReviewer	**100.0**	0.0	**100.0**	0.0	**100.0**	0.0	**100.0**	0.0
In-context prompting — direct LLM review								
GPT-5	**97.5**	2.5	**98.5**	1.5	**97.5**	2.5	**85.5**	10.5
GPT-4.1	**97.0**	3.0	**94.5**	5.5	**97.0**	3.0	**87.0**	5.5
Claude Sonnet 4	**100.0**	0.0	**99.5**	0.5	**100.0**	0.0	**93.5**	2.0
Gemini 2.5 Flash	**70.5**	28.0	**58.5**	41.5	**72.0**	28.0	**47.5**	31.5
DeepSeek-R1	**85.0**	15.0	**88.5**	11.0	**85.0**	15.0	**64.5**	30.5
DeepSeek-V3.1	**94.0**	6.0	**88.0**	12.0	**95.5**	4.5	**84.5**	11.5
Qwen3-32B	**98.0**	2.0	**97.0**	3.0	**98.0**	2.0	**88.5**	11.0
In-context prompting — multi-agent review								
AI-Scientist	**93.0**	7.0	**93.5**	6.5	**93.0**	7.0	**81.5**	17.0
TreeReview	**96.0**	4.0	**96.5**	3.5	**96.0**	4.0	**87.5**	6.5

Table 8: Pairwise evaluation under a second cross-family judge (GPT-5.4).

SQUARE vs.	**Win**	**Tie**	**Lose**	**Win Rate**
DeepReview-7B	200	0	0	100.0
Claude Sonnet 4	193	4	3	96.5
TreeReview	188	4	8	94.0
GPT-4.1	183	4	13	91.5
DeepSeek-V3.1	166	5	29	83.0
AI-Scientist	154	5	41	77.0
Gemini 2.5 Flash	137	5	58	68.5

E RETRIEVAL AUGMENTATION ANALYSIS

Because SQUARE retrieves related papers during the Comparative Novelty Assessment stage, a potential confound is that its gains may partly arise from access to external literature rather than from the structured inquiry workflow itself. To separate these effects, we select four representative

Table 9: Average review length vs. SQuARE's Overall Judgment win rate.

System	Avg. Chars	Win%
SQuARE	8,969	—
DeepReviewer	15,181	100.0
Gemini 2.5 Flash	7,717	89.5
TreeReview	6,540	96.0
Qwen3-32B	6,259	100.0
Claude Sonnet 4	5,432	100.0
AI-Scientist	4,472	79.5
DeepSeek-R1	4,601	93.0
GPT-4.1	4,392	97.5
GPT-5	4,081	98.0
DeepSeek-V3.1	3,801	92.5
CycleReviewer	1,752	100.0

baselines—GPT-4.1, Gemini 2.5 Flash, DeepSeek-R1, and DeepSeek-V3.1—and provide each with the same set of retrieved papers by prepending them to the review prompt (Figure 7). This ensures that all methods are evaluated under comparable information conditions. We then re-evaluate the same 200 submissions, with results reported in Tables 10 and 11.

As shown in both tables, retrieval augmentation leads to only limited improvements for the baselines, which continue to underperform SQuARE on both rule-based and preference-based evaluations. These results suggest that access to related literature alone does not fully account for the observed gains. Instead, the way in which SQuARE incorporates retrieved evidence—through targeted question generation and evidence-grounded novelty comparison—appears to contribute to the improvement in review quality.

Table 10: Rule-based comparison of retrieval-augmented baselines with SQuARE.

Model	Rating		Soundness		Presentation		Contribution	
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
GPT-4.1 + Retrieval	2.13	5.89	1.06	1.36	0.53	0.43	1.09	1.49
Gemini 2.5 Flash + Retrieval	2.10	5.80	0.99	1.28	0.80	0.90	1.16	1.66
DeepSeek-R1 + Retrieval	1.73	4.18	0.84	0.97	0.54	0.48	1.23	1.79
DeepSeek-V3.1 + Retrieval	1.74	4.10	1.03	1.34	0.59	0.52	1.36	2.12
SQuARE (Ours)	**1.03**	**1.67**	**0.42**	**0.29**	**0.47**	**0.35**	**0.50**	**0.37**

Table 11: LLM-as-a-judge win-rate (%) of SQuARE against retrieval-augmented baselines.

Baseline (SQuARE *vs.*)	Overall		Constructive		Analytical		Technical	
	Win↑	Lose↓	Win↑	Lose↓	Win↑	Lose↓	Win↑	Lose↓
GPT-4.1 + Retrieval	**95.5**	4.5	**91.5**	3.0	**96.5**	3.5	**88.0**	3.0
Gemini 2.5 Flash + Retrieval	**75.5**	24.0	**57.0**	29.0	**75.5**	23.0	**61.0**	16.0
DeepSeek-R1 + Retrieval	**89.5**	10.0	**87.0**	7.5	**88.5**	11.5	**74.5**	8.0
DeepSeek-V3.1 + Retrieval	**90.0**	9.5	**82.0**	13.5	**91.5**	8.5	**83.0**	4.5

F Robustness Across Backbone Models

All main experiments use Gemini 2.5 Flash as the backbone. To test whether the gains of SQuARE depend on this particular model, we re-run the full pipeline with DeepSeek-V3.1 on the same 200 papers, while keeping the baselines and Gemini-2.5-Pro judge fixed.

As shown in Table 12, SQuARe achieves a Rating MAE/MSE of $0.98/1.57$ with DeepSeek-V3.1, reducing MAE by 46% over direct DeepSeek-V3.1 generation ($1.83 \rightarrow 0.98$). This is close to the 54% reduction observed with Gemini 2.5 Flash ($2.23 \rightarrow 1.03$), indicating that the gains generalize across backbone models rather than relying on a specific one. The DeepSeek-V3.1 variant is also preferred over all baselines in pairwise evaluation, with win rates ranging from 61.0% to 99.0%, while achieving rating accuracy comparable to the domain-specific fine-tuned DeepReview-7B ($0.98/1.57$ vs. $1.02/1.64$).

Table 12: Results of SQuARe with DeepSeek-V3.1 as the backbone.

Baseline	**MAE ↓**	**MSE ↓**	**Win / Lose**
DeepReview-7B	1.02	1.64	99.0 / 1.0
Claude Sonnet 4	1.05	1.69	94.0 / 6.0
TreeReview	1.37	2.96	87.5 / 12.5
DeepSeek-V3.1 (direct)	1.83	4.40	86.0 / 13.5
AI-Scientist	1.95	5.37	72.5 / 27.5
Gemini 2.5 Flash	2.23	6.31	61.0 / 39.0
GPT-4.1	2.34	6.80	89.5 / 10.5
SQuARe (DeepSeek-V3.1)	**0.98**	**1.57**	—

G Statistical Significance Analysis

To verify the reliability of the reported results, we compute confidence intervals for all main metrics. For rating-based metrics (MAE/MSE), we use bootstrap resampling ($B = 10{,}000$) over the 200 test papers to obtain 95% confidence intervals. For LLM-as-a-judge win rates, we report 95% binomial confidence intervals.

Rating Metrics. Table 13 reports the bootstrap 95% CI for SQuARe. All intervals are narrow, confirming the stability of the reported scores.

Win Rate. Table 14 reports the 95% binomial confidence intervals for Overall Judgment win rates. Across all baselines, the lower bound of the CI exceeds 50%, confirming that SQuARe's superiority is statistically significant in every comparison.

H Case Study

To illustrate how SQuARe's progressive inquiry process translates into concrete review insights, we present a case study on a paper proposing a "Belief State Transformer" (BST) for goal-conditioned sequence generation. Over the course of the review, SQuARe generates 40 questions across 8 sections and the full-paper stage, retrieves 3 related papers via the Comparative Novelty Assessment module, and produces a structured review with detailed strengths, weaknesses, and questions.

Progressive Deepening. During sectional inquiry, SQuARe first asks a high-level question from the abstract (**Q1**: *How does the BST leverage its dual input and next/previous prediction objective to overcome limitations of forward-only transformers?*). Upon reading the introduction, the system spawns a more focused follow-up (**Q1.1.2**: *What is the precise definition of a "belief state," and how is this informal notion formalized within the model?*), which at this stage can only be partially answered: the informal definition is extracted but the formalization is noted as absent. Later, in the theoretical analysis section, the question deepens further (**Q1.1.2.1**: *Given the formal definition of a belief state as a vector v_s that enables sampling from $\Pr(x_{t+1:T}|x_{1:t})$, how does the forward encoding in Theorem 1 achieve this?*), at which point the system traces the reverse-order sampling proof and resolves the earlier gap. This progressive chain—from intuition, to definition, to formal proof—demonstrates how the sectional inquiry stage builds a layered understanding before forming any assessment.

Table 13: Bootstrap 95% confidence intervals for SQuARE rating metrics ($B = 10{,}000$).

Metric	MAE [95% CI]	MSE [95% CI]
Rating	1.03 [0.93, 1.14]	1.67 [1.37, 1.99]
Soundness	0.42 [0.38, 0.47]	0.29 [0.23, 0.35]
Presentation	0.47 [0.42, 0.52]	0.35 [0.29, 0.41]
Contribution	0.50 [0.45, 0.55]	0.37 [0.31, 0.43]

Table 14: Binomial 95% confidence intervals for SQuARE Overall Judgment win rates.

Baseline	Win%	95% CI
CycleReviewer	100.0	[100.0, 100.0]
DeepReviewer	100.0	[100.0, 100.0]
GPT-5	98.0	[96.1, 99.9]
GPT-4.1	97.5	[95.3, 99.7]
Claude Sonnet 4	100.0	[100.0, 100.0]
Gemini 2.5 Flash	89.5	[85.3, 93.7]
DeepSeek-R1	93.0	[89.5, 96.5]
DeepSeek-V3.1	92.5	[88.8, 96.2]
Qwen3-32B	100.0	[100.0, 100.0]
AI-Scientist	79.5	[73.9, 85.1]
TreeReview	96.0	[93.3, 98.7]

Cross-section Synthesis. During the full-paper stage, cross-section inquiry identifies multiple issues that cannot be detected by reading any single section in isolation. **Q21** contrasts the abstract's "domain-independent" claim against the experimental sections, which cover only two synthetic domains: *The paper claims the BST is "domain-independent," yet the empirical evaluation covers only star graphs and TinyStories. The conclusion itself acknowledges these are "proof-of-concept." Can this claim be substantiated?* Similarly, **Q22** challenges the efficiency claim by noting that while default inference requires no extra parameters, the goal-conditioned planning algorithm involves multi-rollout search, yet no FLOPs or latency measurements are reported. These cross-section questions synthesize information scattered across the abstract, method, experiments, and conclusion to surface overstatements and unsupported claims.

Novelty Assessment. The Comparative Novelty Assessment stage retrieves three related papers, including "Backward Learning for Goal-Conditioned Policies" and "Evaluation of LLMs on Syntax-Aware Code Fill-in-the-Middle Tasks." These enable the system to contextualize the BST's bidirectional objective against prior backward-learning and fill-in-the-middle approaches, resulting in a more grounded assessment of the paper's novelty claims.

Outcome. The final review accurately identifies the key weaknesses (*limited domain coverage*, *lack of quantitative efficiency analysis*, *ambiguity in "maximal solution" claims*, *reliance on GPT-4 as a judge*) that align closely with the issues raised by the four human reviewers. For instance, three of the four human reviewers independently flag the narrow evaluation domains, and two raise concerns about scalability—both captured by SQuARE's cross-section synthesis. SQuARE predicts a rating of 6, compared to the human average of 6.25 (individual scores: 6, 5, 8, 6), demonstrating that evidence-grounded inquiry yields reviews that are both well-calibrated in scoring and substantively aligned with expert judgment.

I Prompts Used in SQuARE

We provide the complete prompts used throughout the SQuARE framework. The Sectional Inquiry stage includes prompts for initial Q&A generation (Figure 8), section-wise question generation (Figure 9), and section-wise question answering (Figure 10). The Cross-section Inquiry stage contains prompts for cross-section question generation (Figure 11) and answering (Figure 12). The Comparative Novelty Assessment stage includes four prompts for search query generation (Figure 13), paper relevance scoring (Figure 14), novelty question generation (Figure 15), and novelty question

answering (Figure 16). Finally, the Review Synthesis stage uses a novelty-conditioned synthesis prompt (Figure 17), together with a fallback variant for cases in which no relevant papers are available (Figure 18). The prompt used for baseline review models is shown in Figure 19.

(a) System Prompt

```
---***---
SYSTEM PROMPT
---***---

You are a neutral arbitrator evaluating peer review comments for academic papers.
Your role is to compare reviews and determine which is better on each aspect. Your
judgments must be strictly based on verifiable evidence from the paper and reviews.

For each evaluation, you must:

1. Thoroughly understand the paper by analyzing:
   - Research objectives and contributions
   - Methodology and experiments
   - Claims and evidence
   - Results and conclusions

2. For each review, methodically examine:
   - Claims made about the paper
   - Evidence cited to support claims
   - Technical assessments and critiques
   - Suggested improvements

You will evaluate reviews based on these key aspects:

**Technical Accuracy**
- Are claims consistent with paper content?
- Is evidence properly interpreted?
- Are technical assessments valid?
- Are critiques well-supported?

**Constructive Value**
- How actionable is the feedback?
- Are suggestions specific and feasible?
- Is criticism balanced with strengths?
- Would authors understand how to improve?

**Analytical Depth**
- How thoroughly are key aspects examined?
- Is analysis appropriately detailed?
- Are important elements addressed?
- Is assessment comprehensive?

**Input Format:**
- Complete paper text
- Assistant A's review
- Assistant B's review

**Output Format:**

For each aspect, only provide a judgment:

```
**Technical Accuracy:** **Better Assistant:** [A or B or Tie]
**Constructive Value:** **Better Assistant:** [A or B or Tie]
**Analytical Depth:** **Better Assistant:** [A or B or Tie]
```

Conclude with:

```
**Overall Judgment:**
**Better Assistant:** [A or B or Tie]
```

Key Requirements:
1. Base all judgments on concrete evidence from the paper and reviews
2. Maintain neutral arbitrator perspective
3. Judge Tie only when both reviews demonstrate truly comparable quality
4. Do NOT provide analysis, reasoning, or explanations - only output the judgments

Begin analysis after receiving complete materials.
```

(b) User Prompt Template (randomized A/B order)

```
# Paper:
{paper_content}

---***---
---***---
---***---
#Assistant A:
{assistant_a_review}

---***---
---***---
---***---
#Assistant B:
{assistant_b_review}
```

Figure 6: System prompt and user prompt template used for the LLM-as-a-judge win-rate evaluation. The order of Assistant A and Assistant B is randomized for each evaluation instance to mitigate position bias.

Prompt for Retrieval-Augmented Baseline Review

```
You are an expert scientific reviewer. Your task is to write a thorough, critical, and
constructive review for a paper submitted for publication, evaluating whether it brings
sufficient value to the community and contributes new knowledge.

To help you better assess this paper, we provide a list of related papers retrieved from
the literature. You should use these related works to:
- Evaluate the novelty and originality of the submitted paper
- Assess whether the paper adequately discusses and differentiates itself from prior work
- Identify any missing references or comparisons
- Better understand the context and state-of-the-art in the relevant research area

REVIEW CRITERIA
Your review must include the following sections:

Summary: [Provide a concise summary of the paper, highlighting its main objectives,
methodology, results, and conclusions.]
Strengths and Weaknesses: [Critically analyze the strengths and weaknesses of the paper.
Consider the significance of the research question, the robustness of the methodology, and
the relevance of the findings. Leverage the related papers to assess novelty and
positioning.]
Questions: [Please list up and carefully describe any questions and suggestions for the
authors.]
Soundness: [1 poor | 2 fair | 3 good | 4 excellent]
Presentation: [1 poor | 2 fair | 3 good | 4 excellent]
Contribution: [1 poor | 2 fair | 3 good | 4 excellent]
Flag for Ethics Review: [YES or NO]
Rating: [1 strong reject | 2 reject | 3 reject, not good enough | 4 possibly reject | 5
marginally below | 6 marginally above | 7 accept, minor improvements | 8 accept, good
paper | 9 strong accept | 10 strong accept, highlighted]
Confidence: [1 educated guess | 2 willing to defend | 3 fairly confident | 4 confident | 5
absolutely certain]

INPUT
- Paper Content: {paper_content}
- Related Papers: {related_papers}

OUTPUT FORMAT
**Summary:** Summary content
**Strengths:** Strengths result
**Weaknesses:** Weaknesses result
**Questions:** Questions result
**Soundness:** Soundness result
**Presentation:** Presentation result
**Contribution:** Contribution result
**Rating:** Rating result
**Confidence:** Confidence result
```

Figure 7: Prompt used for the retrieval-augmented baselines in Section E. The retrieved related papers are prepended to the standard review prompt.

Prompt for Initial Q&A (Sectional Inquiry)

```
You are a professional academic reviewer. Your task is to generate questions that
help further understand the content of a paper based on its title and abstract.
The questions should capture the key aspects of the work and follow the MECE
principle—mutually exclusive and collectively exhaustive.

TASK REQUIREMENTS:

1. Mutually Exclusive: Each question should be distinct from the others, addressing
different aspects of the paper.
2. Collectively Exhaustive: The questions should collectively cover all the
important aspects of the paper, ensuring no important information is missed.
3. Importance: The questions should focus on concepts, assumptions, or results
essential to the paper's core contributions, avoiding trivial or peripheral details.
3. Relevance: The questions should be directly related to the paper's content and
its research contributions.
4. Clarity: The questions should be clear and easy to understand, with no ambiguity
or vagueness.
5. Conciseness: The questions should be concise and to the point, with no
unnecessary words or phrases.
6. No more than 5 questions in total.

Input:
Title: <TITLE>{title}</TITLE>
Abstract: <ABSTRACT>{abstract}</ABSTRACT>

Output Format:
[1] write the question 1 here
[2] write the question 2 here
...
```

Figure 8: Prompt for generating initial questions from the paper's title and abstract (Sectional Inquiry — Initial Q&A).

Prompt for Section-wise Question Generation (Sectional Inquiry)

```
You are a professional academic reviewer evaluating a manuscript under review and
recording notes in a question-and-answer format while reading.
Based on the section currently being read, please formulate additional questions
that help further deepen the understanding of the paper or highlight aspects that
are unreasonable or do not make sense.
These questions may either be new questions arising from the current section or
refined or newly inspired questions prompted by this section in relation to earlier
questions.

TASK REQUIREMENTS:
1. Relevance: The questions should be directly related to the paper's content and
its research contributions.
2. Clarity: The questions should be clear and easy to understand, with no ambiguity
or vagueness.
3. Conciseness: The questions should be concise and to the point, with no
unnecessary words or phrases.
4. Importance: The questions should focus on concepts, assumptions, or results
essential to the paper's core contributions, or on critical writing issues, avoiding
 trivial or peripheral details."
5. Self-containment: Newly proposed questions should be independent and complete,
avoiding pronouns or implicit references. Each question should contain all necessary
 context explicitly.
6. Different sections of the paper require attention to different aspects.
    For example:
    - The Introduction should be evaluated for clarity of motivation, problem
    formulation, and main contributions.
    - The Related Work section should be assessed for coverage, positioning, and
    differentiation from prior studies.
    - The Methodology section should be examined for soundness, clarity, and
    completeness of the proposed approach.
    - The Experiments section should be reviewed for rigor of design,
    appropriateness of datasets and metrics, fairness of comparisons, and support
    for the claims.
    - The Analysis / Discussion section should be considered for the depth of
    interpretation, articulation of strengths and limitations, and discussion of
    broader implications.
7. Clearly label question numbers: Use a consistent and standardized numbering
system, and rely on punctuation to indicate the relationships between questions.
    For example:
    - [1.1] denotes the first sub-question of Question [1],
    - [2] denotes a new independent question relative to Question [1],
    - [1.1.1] denotes a further sub-question under [1.1].
    This helps clearly convey the structure and logical hierarchy of the questions,
    making them easier to understand and reference.
8. Newly proposed questions must not duplicate existing question numbers.
9. No more than 5 questions in total.

Input:
paper metadata: <TITLE>{title}</TITLE> <ABSTRACT>{abstract}</ABSTRACT>
previous questions and answers:
<QUESTIONS_AND_ANSWERS>{questions_and_answers}</QUESTIONS_AND_ANSWERS>
section content: <SECTION>{section}</SECTION>

Output Format:
[question_number]write the question here
[question_number]write the question here
...
```

Figure 9: Prompt for generating new questions while reading each section (Sectional Inquiry — Section-wise Question Generation).

Prompt for Section-wise Question Answering (Sectional Inquiry)

```
You are a professional academic reviewer evaluating a manuscript under review and
recording notes in a question-and-answer format while reading.
Based on the content of the current section, identify any information that is
relevant to previously unresolved questions and, where applicable, provide
corresponding answers in your own words.
If the current section does not contain information pertinent to a given question,
explicitly note that the question remains unresolved and should be addressed in
subsequent sections (Note: "unresolved").

TASK REQUIREMENTS:
1.Clarity: Answers should be clear, precise, and free of ambiguity.
2.Conciseness: Answers should be succinct and focused, avoiding unnecessary wording.
3.Answer according to the original text; do not make up information.
4.Fallback Rule: If no meaningful notes can be produced based on the available
content, simply write "[number]unresolved". This option is encouraged when
appropriate.
5.Each question is assigned a numbered label (e.g., [1], [1.1]), answer according to
 the question number.
6.You are required to provide responses only for the unresolved questions. Don't
respond to the questions that have been resolved.
7.Reasoning: Include your reasoning process(chain-of-thought, cot) in each answer to
 ensure responses are accurate, complete, and free from errors or omissions.

Input:
Paper metadata: <TITLE>{title}</TITLE> <ABSTRACT>{abstract}</ABSTRACT>

previous questions and answers:
<QUESTIONS_AND_ANSWERS>{questions_and_answers}</QUESTIONS_AND_ANSWERS>

section content: <SECTION>{section}</SECTION>

Output Format:
[number] "write the answer with cot here"
[number] "write the answer with cot here"
...
```

Figure 10: Prompt for answering questions based on the current section content (Sectional Inquiry — Section-wise Question Answering).

Prompt for Cross-section Question Generation (Cross-section Inquiry)

```
You are a professional academic reviewer evaluating a manuscript under review and
recording notes in a question-and-answer format while reading.
You now have access to the full text of the paper. Based on the complete paper
content along with your previous Q&A notes, please identify and formulate additional
 questions that have not been adequately addressed or explored in the existing
questions.
These questions should focus on identifying problems, flaws, limitations, or
critical issues that may have been overlooked in previous analysis.

TASK REQUIREMENTS:
1. Gap Analysis: Carefully review all previous questions and answers to identify
aspects of the paper that have not been thoroughly examined or questioned.
2. Problem-Focused: The questions should prioritize identifying problems, defects,
limitations, methodological flaws, logical inconsistencies, or areas where the
paper's claims are not well-supported.
3. Critical Thinking: Focus on weaknesses that could significantly impact the
paper's validity, reliability, or contribution, rather than minor issues.
4. Relevance: The questions should be directly related to the paper's content,
methodology, experiments, or claims.
5. Clarity: The questions should be clear and easy to understand, with no ambiguity
or vagueness.
6. Conciseness: The questions should be concise and to the point, with no
unnecessary words or phrases.
7. Self-containment: Newly proposed questions should be independent and complete,
avoiding pronouns or implicit references. Each question should contain all necessary
 context explicitly.
8. Clearly label question numbers: Use a consistent and standardized numbering
system, and rely on punctuation to indicate the relationships between questions.
    For example:
    - [1.1] denotes the first sub-question of Question [1],
    - [2] denotes a new independent question relative to Question [1],
    - [1.1.1] denotes a further sub-question under [1.1].
    This helps clearly convey the structure and logical hierarchy of the questions,
    making them easier to understand and reference.
9. Newly proposed questions must not duplicate existing question numbers.
10. No more than 5 questions in total.

Areas to Consider (but not limited to):
- Methodological flaws or limitations in the proposed approach
- Issues with experimental design, dataset selection, or evaluation metrics
- Unsupported or overstated claims
- Missing comparisons with relevant baselines or prior work
- Logical inconsistencies or contradictions in the paper
- Potential issues with reproducibility or generalizability
- Critical missing information or unclear explanations
- Problems with the theoretical foundation or assumptions

Input:
Paper full text: <FULL_PAPER>{full_paper}</FULL_PAPER>

previous questions and answers:
<QUESTIONS_AND_ANSWERS>{questions_and_answers}</QUESTIONS_AND_ANSWERS>

Output Format:
[question_number]write the question here
[question_number]write the question here
...
```

Figure 11: Prompt for generating cross-section questions based on the full paper (Cross-section Inquiry — Question Generation).

Prompt for Cross-section Question Answering (Cross-section Inquiry)

```
You are a professional academic reviewer evaluating a manuscript under review and
recording notes in a question-and-answer format while reading.

You now have access to the full text of the paper. Based on the complete paper
content along with your previous Q&A notes, please answer all questions that remain
unanswered.
For questions that still cannot be resolved after reading the entire paper, list the
 available information gathered so far and explicitly explain why the question
cannot be adequately answered.

TASK REQUIREMENTS:
1.Clarity: Answers should be clear, precise, and free of ambiguity.
2.Conciseness: Answers should be succinct and focused, avoiding unnecessary wording.
3.Answer according to the original text; do not make up information.
4.Each question is assigned a numbered label (e.g., [1], [1.1]), answer according to
 the question number.
5.You are required to provide responses only for the unresolved questions. Don't
respond to the questions that have been resolved.
6.Reasoning: Include your reasoning process (chain-of-thought, cot) in each answer
to ensure responses are accurate, complete, and free from errors or omissions.

Input:
Paper full text: <FULL_PAPER>{full_paper}</FULL_PAPER>

previous questions and answers:
<QUESTIONS_AND_ANSWERS>{questions_and_answers}</QUESTIONS_AND_ANSWERS>

Output Format:
[number] write the answer with cot here
[number] write the answer with cot here
...
```

Figure 12: Prompt for answering all remaining questions with access to the full paper (Cross-section Inquiry — Question Answering).

Prompt for Search Query Generation (Comparative Novelty Assessment)

```
You are now an academic paper review expert. Your task is to generate three short
search queries based on the content of the paper.
Here is a research paper:

<PAPER_BEGIN>
{paper}
<PAPER_END>

Task:
Please generate three short search queries based on the content of the paper to
retrieve related literature and enrich background knowledge for the subsequent
review.
Each query should be relevant to the paper, kept as concise as possible, independent
 from one another, and focused on broader research aspects.
Respond in the following format:

<Query> generated query 1 here </Query>
<Query> generated query 2 here </Query>
<Query> generated query 3 here </Query>

Begin your response with the queries only, without any additional text.
```

Figure 13: Prompt for generating search queries to retrieve related papers (Comparative Novelty Assessment — Query Generation).

Prompt for Paper Relevance Scoring (Comparative Novelty Assessment)

```
You are now an academic paper review expert. Your task is to score the relatedness
of a paper to the object paper.

Task:
Given a paper's metadata (title, abstract), please score the relatedness of the
paper to the object paper.
The score should be a number between 0 and 100, where 0 means not related and 100
means very related.
Respond in the following format:

<Score> score here </Score>

Input:
- Object paper metadata: {object_paper_metadata}
- Paper to check metadata: {paper_metadata}

Directly output the result, without any additional text.
```

Figure 14: Prompt for scoring the relevance of a candidate paper to the manuscript under review (Comparative Novelty Assessment — Paper Relevance Scoring).

Prompt for Novelty Question Generation (Comparative Novelty Assessment)

```
You are an experienced and rigorous academic peer reviewer with deep familiarity in
this research field and its existing literature.

I will provide you with:
1. The Title and Abstract of a target paper under review
2. The Titles and Abstracts of several related papers

Your task is to evaluate the novelty of the target paper and formulate critical
reviewer questions focused specifically on whether the paper makes a clear,
substantive, and defensible novel contribution beyond existing work.

Review Guidelines:

1. Comparison with Prior Work
- Explicitly consider the provided related papers
- Identify potential overlaps in methodology, problem formulation, assumptions, or
conclusions
- Question whether the differences claimed by the authors are clearly articulated
and genuinely meaningful

2. Focus on True Novelty
- Avoid vague or generic statements (e.g., "the novelty is unclear")
- Probe whether the proposed contributions are:
  - Minor variations, parameter changes, or combinations of existing methods
  - Straightforward extensions or applications of known approaches
  - Implicitly or explicitly covered by the related papers

3. Emphasize Substantive Contribution over Presentation
- Distinguish between conceptual novelty, methodological novelty, problem-setting
novelty, and engineering or implementation-level improvements
- Question whether any claimed contributions amount to re-labeling, re-framing, or
re-packaging of prior work

4. Question-Only Output
- Do not provide conclusions or judgments
- Output only reviewer-style questions that could reasonably be asked during peer
review

Output Format:

- Overall Novelty Questions (no more than 3 items)
  (High-level questions assessing whether the paper provides an independent and
  meaningful contribution)
  respond in the following format:
  <Overall Novelty Questions> write the question here </Overall Novelty Questions>
  <Overall Novelty Questions> write the question here </Overall Novelty Questions>

- Comparison-to-Prior-Work Questions
  (Each question should explicitly reference one or more related papers by their
  Title)
  respond in the following format:
  <Comparison-to-Prior-Work Questions> write the question here
  </Comparison-to-Prior-Work Questions>
  <Comparison-to-Prior-Work Questions> write the question here
  </Comparison-to-Prior-Work Questions>

- Questions on Potential Incremental or Reused Contributions (1–2 items)
  respond in the following format:
  <Questions on Potential Incremental or Reused Contributions> write the question
  here </Questions on Potential Incremental or Reused Contributions>
  <Questions on Potential Incremental or Reused Contributions> write the question
  here </Questions on Potential Incremental or Reused Contributions>

The questions should be written in a professional, formal peer-review tone and be
specific, pointed, and technically grounded.

Content to Analyze:

[Target Paper]
Title: {target_paper_title}
Abstract: {target_paper_abstract}

[Related Paper List]
{related_papers_content}
```

Figure 15: Prompt for generating novelty-oriented questions by comparing the manuscript with related papers (Comparative Novelty Assessment — Question Generation).

Prompt for Novelty Question Answering (Comparative Novelty Assessment)

```
You are an experienced and rigorous academic peer reviewer with deep familiarity in
this research field and its existing literature.

Previously, you reviewed the title and abstract of a target paper along with related
 papers, and generated a set of critical questions regarding the paper's novelty and
 contributions.
Now, you are provided with the full content of both the target paper and the related
 papers. Your task is to answer those questions based on a thorough analysis of the
complete paper contents.

I will provide you with:
1. The full content of the target paper under review
2. The abstract of related papers
3. The previously generated novelty questions

Answer Guidelines:

1. Evidence-Based Responses
- Base your answers strictly on the content provided in the papers
- Cite specific sections, methods, results, or claims from the target paper when
answering
- Reference concrete details from related papers when making comparisons

2. Objective and Balanced Assessment
- Provide fair and balanced answers that acknowledge both strengths and weaknesses
- Avoid overly harsh criticism or unwarranted praise
- If the paper adequately addresses a concern, state so clearly with supporting
evidence

3. Structured and Specific Answers
- Answer each question directly and specifically
- Avoid vague or generic responses
- If a question cannot be fully answered based on the provided content, explicitly
state what information is missing

4. Technical Rigor
- Ensure technical accuracy in your assessments
- Compare methodological details, experimental setups, and theoretical contributions
 precisely
- Identify subtle but meaningful differences or similarities between works

Output Format:

For each question, respond in the following format:

<Question> [Restate the original question here] </Question>
<Answer> [Provide your detailed, evidence-based answer here] </Answer>

Content to Analyze:

[Target Paper - Full Content]
{target_paper_content}

[Related Papers]
{related_papers_content}

[Previously Generated Questions]
{innovation_questions_content}


Directly output the result, without any additional text.
```

Figure 16: Prompt for answering novelty questions based on the full paper and related work (Comparative Novelty Assessment — Question Answering).

Prompt for Review Synthesis (with Novelty Assessment)

```
You are an expert scientific reviewer. Your task is
to write a thorough, critical, and constructive
review for a paper submitted for publication,
evaluating whether it brings sufficient value to the
 community and contributes new knowledge.

Before writing the formal review, you have:
1. Reviewed related papers in the field to
understand the research landscape.
2. Conducted a preliminary novelty assessment by
generating and answering critical questions about
the paper's innovation compared to prior work.
3. Carefully read the entire paper section by
section, and made notes in the form of Q&A pairs for
 each section.

Now, based on all the provided materials, please
write a detailed and comprehensive review.

**Task Requirements**:
1. Ensure your review is comprehensive and
substantive, covering all important aspects of the
paper. Present your review with a clear and
well-organized structure, using bullet points for
readability.
2. The input contains:
    - The full paper text, with each section
    followed by relevant Q&A pairs (your reading
    notes).
    - Abstracts of related papers in the field for
    context.
    - Innovation Q&A pairs that assess the paper's
    novelty compared to prior work.
3. When writing the Strengths section, briefly
identify the paper's genuine contributions.
4. When writing the Weaknesses section, you should:
    - First focus on the paper's content itself to
    develop your own understanding, then identify
    its inherent weaknesses and flaws based on your
    independent analysis.
    - Leverage the Innovation Q&A pairs to identify
    issues with novelty claims, incremental
    contributions, or overlap with existing work.
    - Use the related papers' abstracts to assess
    whether the paper adequately differentiates
    itself from prior work and properly acknowledges
     relevant literature.
    - Critically evaluate the section-level Q&A
    pairs. Only incorporate points that reveal
    significant flaws or limitations.
    - Do not simply restate the Q&A pairs. Instead,
    synthesize them to provide a high-level critique
     of the paper's methodology, experiments, and
    claims.
    - Ensure your Weaknesses section is thorough,
    well-organized, and focuses on substantive
    issues that affect the paper's quality and
    validity.
5. When writing the Questions section, you should:
    - Focus on points where author clarification
    could change your assessment or resolve
    ambiguities in the paper.
    - Ask about unclear assumptions, missing
    details, or potential concerns that the authors
    may be able to address.
    - Include constructive suggestions that could
    help improve the paper if addressed.
    - Avoid restating weaknesses as questions;
    instead, focus on genuine inquiries and
    actionable recommendations.
    - Prioritize questions that target the most
    critical aspects of the work.
6. When assigning scores (Soundness, Presentation,
Contribution, Rating), carefully consider:
    - The Innovation Q&A analysis results when
    evaluating the Contribution score.
    - The comparison with related papers when
    assessing the significance and novelty of the
    work.

## REVIEW CRITERIA

Your review must include the following sections:

Summary: [Provide a concise summary of the paper,
highlighting its main objectives, methodology,
results, and conclusions.]
Strengths and Weaknesses: [Critically analyze the
strengths and weaknesses of the paper. Consider the
significance of the research question, the
robustness of the methodology, the relevance of the
findings, and the novelty compared to related work.]
Questions: [Please list up and carefully describe
any questions and suggestions for the authors. Think
 of the things where a response from the author can
change your opinion, clarify a confusion or address
a limitation.]
Soundness: [Please assign the paper a numerical
rating on the following scale to indicate the
soundness of the technical claims, experimental and
research methodology and on whether the central
claims of the paper are adequately supported with
evidence. You are only allowed to choose from the
following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Presentation: [Please assign the paper a numerical
rating on the following scale to indicate the
quality of the presentation. This should take into
account the writing style and clarity, as well as
contextualization relative to prior work. You are
only allowed to choose from the following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Contribution: [Please assign the paper a numerical
rating on the following scale to indicate the
quality of the overall contribution this paper makes
 to the research area being studied. Consider the
Innovation Q&A analysis when making this assessment.
 You are only allowed to choose from the following
options:
    1 poor
    2 fair
    3 good
    4 excellent]
Flag for Ethics Review: Indicate whether the paper
should undergo an ethics review [YES or NO].
Rating: [Give this paper an appropriate rating. You
are only allowed to choose from the following
options:
    1 strong reject
    2 reject, significant issues present
    3 reject, not good enough
    4 possibly reject, but has redeeming facets
    5 marginally below the acceptance threshold
    6 marginally above the acceptance threshold
    7 accept, but needs minor improvements
    8 accept, good paper
    9 strong accept, excellent work
    10 strong accept, should be highlighted at the
    conference]
Confidence: [Rate your confidence level in your
assessment, you are only allowed to choose from the
following options:
    1 Your assessment is an educated guess.
    2 You are willing to defend your assessment, but
     it is quite likely that you did not understand
    the central parts of the submission or that you
    are unfamiliar with some pieces of related work.
    3 You are fairly confident in your assessment.
    4 You are confident in your assessment, but not
    absolutely certain.
    5 You are absolutely certain about your
    assessment.

## INPUT

### Paper Content with Section-Level Q&A Pairs:
{paper_content}

### Related Papers (Title and Abstract):
{related_papers_content}

### Innovation Assessment Q&A:
{innovation_qa_content}

## OUTPUT FORMAT
Here is the template for a review format. You must
follow this format to output the integrated review
results:
**Summary:**
Summary content
**Strengths:**
Strengths result
**Weaknesses:**
Weaknesses result
**Questions:**
Questions result
**Soundness:**
Soundness result
**Presentation:**
Presentation result
**Contribution:**
Contribution result
**Rating:**
Rating result
**Confidence:**
Confidence result

Your final review, do not include any additional
commentary:
```

Figure 17: Prompt for synthesizing the final review from the inquiry notes, related papers, and novelty assessment (Review Synthesis — with Novelty Assessment).

Prompt for Review Synthesis (without Related Papers)

```
You are an expert scientific reviewer. Your task is
to write a thorough, critical, and constructive
review for a paper submitted for publication,
evaluating whether it brings sufficient value to the
 community and contributes new knowledge.

Before writing the formal review, you have carefully
 read the entire paper section by section and made
notes in the form of Q&A pairs for each section.

Now, based on all the provided materials, please
write a detailed and comprehensive review.

**Task Requirements**:
1. Ensure your review is comprehensive and
substantive, covering all important aspects of the
paper. Present your review with a clear and
well-organized structure, using bullet points for
readability.
2. The input contains the full paper text, with each
 section followed by relevant Q&A pairs (your
reading notes).
3. When writing the Strengths section, briefly
identify the paper's genuine contributions.
4. When writing the Weaknesses section, you should:
    - Focus on the paper's content itself to develop
     your own understanding, then identify its
     inherent weaknesses and flaws based on your
     independent analysis.
    - Critically evaluate the section-level Q&A
     pairs. Only incorporate points that reveal
     significant flaws or limitations.
    - Do not simply restate the Q&A pairs. Instead,
     synthesize them to provide a high-level critique
      of the paper's methodology, experiments, and
     claims.
    - Ensure your Weaknesses section is thorough,
     well-organized, and focuses on substantive
     issues that affect the paper's quality and
     validity.
5. When writing the Questions section, you should:
    - Focus on points where author clarification
     could change your assessment or resolve
     ambiguities in the paper.
    - Ask about unclear assumptions, missing
     details, or potential concerns that the authors
     may be able to address.
    - Include constructive suggestions that could
     help improve the paper if addressed.
    - Avoid restating weaknesses as questions;
     instead, focus on genuine inquiries and
     actionable recommendations.
    - Prioritize questions that target the most
     critical aspects of the work.
6. When assigning scores (Soundness, Presentation,
Contribution, Rating), carefully consider:
    - Soundness: Evaluate whether the technical
     claims are well-supported by theoretical
     reasoning or experimental evidence. Consider the
      rigor of proofs, validity of assumptions, and
     reproducibility of results.
    - Presentation: Assess the clarity of writing,
     logical flow of arguments, quality of
     figures/tables, and how well the paper
     contextualizes itself relative to prior work.
    - Contribution: Judge the novelty and
     significance of the work. Consider whether it
     introduces new ideas, methods, or insights that
     meaningfully advance the field beyond
     incremental improvements.
    - Rating: Derive the overall rating by weighing
     all factors holistically -- a paper with strong
     contributions but notable methodological flaws,
     or vice versa, should be rated accordingly.
     Ensure your rating is consistent with the
     severity and breadth of the weaknesses you
     identified.

## REVIEW CRITERIA

Your review must include the following sections:

Summary: [Provide a concise summary of the paper,
highlighting its main objectives, methodology,
results, and conclusions.]
Strengths and Weaknesses: [Critically analyze the
strengths and weaknesses of the paper. Consider the
significance of the research question, the
robustness of the methodology, the relevance of the
findings, and the novelty compared to related work.]
Questions: [Please list up and carefully describe
any questions and suggestions for the authors. Think
 of the things where a response from the author can
change your opinion, clarify a confusion or address
a limitation.]
Soundness: [Please assign the paper a numerical
rating on the following scale to indicate the
soundness of the technical claims, experimental and
research methodology and on whether the central
claims of the paper are adequately supported with
evidence. You are only allowed to choose from the
following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Presentation: [Please assign the paper a numerical
rating on the following scale to indicate the
quality of the presentation. This should take into
account the writing style and clarity, as well as
contextualization relative to prior work. You are
only allowed to choose from the following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Contribution: [Please assign the paper a numerical
rating on the following scale to indicate the
quality of the overall contribution this paper makes
 to the research area being studied. Consider the
Innovation Q&A analysis when making this assessment.
 You are only allowed to choose from the following
options:
    1 poor
    2 fair
    3 good
    4 excellent]
Flag for Ethics Review: Indicate whether the paper
should undergo an ethics review [YES or NO].
Rating: [Give this paper an appropriate rating. You
are only allowed to choose from the following
options:
    1 strong reject
    2 reject, significant issues present
    3 reject, not good enough
    4 possibly reject, but has redeeming facets
    5 marginally below the acceptance threshold
    6 marginally above the acceptance threshold
    7 accept, but needs minor improvements
    8 accept, good paper
    9 strong accept, excellent work
    10 strong accept, should be highlighted at the
    conference]
Confidence: [Rate your confidence level in your
assessment, you are only allowed to choose from the
following options:
    1 Your assessment is an educated guess.
    2 You are willing to defend your assessment, but
     it is quite likely that you did not understand
    the central parts of the submission or that you
    are unfamiliar with some pieces of related work.
    3 You are fairly confident in your assessment.
    4 You are confident in your assessment, but not
    absolutely certain.
    5 You are absolutely certain about your
    assessment.

## INPUT

### Paper Content with Section-Level Q&A Pairs:
{paper_content}

## OUTPUT FORMAT
Here is the template for a review format. You must
follow this format to output the integrated review
results:
**Summary:**
Summary content
**Strengths:**
Strengths result
**Weaknesses:**
Weaknesses result
**Questions:**
Questions result
**Soundness:**
Soundness result
**Presentation:**
Presentation result
**Contribution:**
Contribution result
**Rating:**
Rating result
**Confidence:**
Confidence result

Your final review, do not include any additional
commentary:
```

Figure 18: Prompt for synthesizing the final review when related papers are unavailable (Review Synthesis — without Related Papers).

Prompt for Direct LLM Review Baseline

```
You are an expert scientific reviewer.
 Your task is to write a thorough,
critical, and constructive review for
a paper submitted for publication,
evaluating whether it brings
sufficient value to the community and
contributes new knowledge.

## REVIEW CRITERIA

Your review must include the following
 sections:

Summary: [Provide a concise summary of
 the paper, highlighting its main
objectives, methodology, results, and
conclusions.]
Strengths and Weaknesses: [Critically
analyze the strengths and weaknesses
of the paper. Consider the
significance of the research question,
 the robustness of the methodology,
and the relevance of the findings.]
Questions: [Please list up and
carefully describe any questions and
suggestions for the authors. Think of
the things where a response from the
author can change your opinion,
clarify a confusion or address a
limitation.]
Soundness: [Please assign the paper a
numerical rating on the following
scale to indicate the soundness of the
 technical claims, experimental and
research methodology and on whether
the central claims of the paper are
adequately supported with evidence.
You are only allowed to choose from
the following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Presentation: [Please assign the paper
 a numerical rating on the following
scale to indicate the quality of the
presentation. This should take into
account the writing style and clarity,
 as well as contextualization relative
 to prior work. You are only allowed
to choose from the following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Contribution: [Please assign the paper
 a numerical rating on the following
scale to indicate the quality of the
overall contribution this paper makes
to the research area being studied.
You are only allowed to choose from
the following options:
    1 poor
    2 fair
    3 good
    4 excellent]
Flag for Ethics Review: Indicate
whether the paper should undergo an
ethics review [YES or NO].

Rating: [Give this paper an
appropriate rating. You are only
allowed to choose from the following
options:
    1 strong reject
    2 reject, significant issues
    present
    3 reject, not good enough
    4 possibly reject, but has
    redeeming facets
    5 marginally below the acceptance
    threshold
    6 marginally above the acceptance
    threshold
    7 accept, but needs minor
    improvements
    8 accept, good paper
    9 strong accept, excellent work
    10 strong accept, should be
    highlighted at the conference]
Confidence: [Rate your confidence
level in your assessment, you are only
 allowed to choose from the following
options:
    1 Your assessment is an educated
    guess.
    2 You are willing to defend your
    assessment, but it is quite likely
     that you did not understand the
    central parts of the submission or
     that you are unfamiliar with some
     pieces of related work.
    3 You are fairly confident in your
     assessment.
    4 You are confident in your
    assessment, but not absolutely
    certain.
    5 You are absolutely certain about
     your assessment.

INPUT
- Paper Content: {paper_content}

OUTPUT FORMAT
Here is the template for a review
format. You must follow this format to
 output the integrated review results:
**Summary:**
Summary content
**Strengths:**
Strengths result
**Weaknesses:**
Weaknesses result
**Questions:**
Questions result
**Soundness:**
Soundness result
**Presentation:**
Presentation result
**Contribution:**
Contribution result
**Rating:**
Rating result
**Confidence:**
Confidence result

Your final review, do not include any
additional commentary:
```

Figure 19: Prompt used for all direct LLM review baselines.