Automatic Paper Reviewing with Heterogeneous Graph Reasoning over LLM-Simulated Reviewer-Author Debates

Shuaimin Li [1, *], **Liyang Fan**[2, 3, *], **Yufang Lin**[4, *], **Zeyang Li**[5], **Xian Wei**[4], **Shiwen Ni** [3, †]
Hamid Alinejad-Rokny [6], **Min Yang** [1, 3, †]

[1]Shenzhen Key Laboratory for High Performance Data Mining, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[3] Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology, Shenzhen, China
[4] East China Normal University, Shanghai, China
[5] University of Science and Technology of China, Suzhou, China
[6] University of New South Wales, Sydney, New South Wales, Australia
sm.li2@siat.ac.cn, 2410833006@mails.szu.edu.cn, 71265902002@stu.ecnu.edu.cn, zeyangli@163.com,
xian.wei@tum.de, sw.ni@siat.ac.cn,h.alinejad@unsw.edu.au, min.yang@siat.ac.cn

Abstract

Existing paper review methods often rely on superficial manuscript features or directly on large language models (LLMs), which are prone to hallucinations, biased scoring, and limited reasoning capabilities. Moreover, these methods often fail to capture the complex argumentative reasoning and negotiation dynamics inherent in reviewer-author interactions. To address these limitations, we propose **ReViewGraph** (Reviewer-Author Debates Graph Reasoner), a novel framework that performs heterogeneous graph reasoning over LLM-simulated multi-round reviewer-author debates. In our approach, reviewer-author exchanges are simulated through LLM-based multi-agent collaboration. Diverse opinion relations (e.g., acceptance, rejection, clarification, and compromise) are then explicitly extracted and encoded as typed edges within a heterogeneous interaction graph. By applying graph neural networks to reason over these structured debate graphs, ReViewGraph captures fine-grained argumentative dynamics and enables more informed review decisions. Extensive experiments on three datasets demonstrate that ReViewGraph outperforms strong baselines with an average relative improvement of 15.73%, underscoring the value of modeling detailed reviewer–author debate structures.

Code — https://github.com/relic-yuexi/ReViewGraph

Introduction

Peer review is essential to scientific progress, ensuring the quality, validity, and originality of research (Alberts, Hanson, and Kelner 2008). High-quality reviews help guide the research community toward impactful contributions. However, the recent surge in paper submissions across disciplines (Drozdz and Ladomery 2024) has placed increasing strain on the peer review process. This exponential growth imposes a significant burden on human reviewers, making it increasingly difficult to ensure timely and consistent evaluations (Stelmakh et al. 2021). Moreover, peer review is inherently subjective. In some cases, reviewers may be careless, biased, or even malicious, thereby undermining the fairness and reliability of the decision-making process (Stelmakh et al. 2021; Zhang et al. 2022). These challenges have motivated the development of automatic reviewing systems as a promising solution to alleviate reviewer workload and promote more objective evaluations at scale.

With the rapid advancement of large language models (LLMs), the field of automatic peer review has made notable progress in recent years. In the computer science community, several top-tier conferences have begun experimenting with LLM-based reviewing systems as supplementary tools to assist human reviewers. For instance, ICLR 2025 introduced an LLM-powered automatic review system to complement human assessments[1], while AAAI 2026 has officially announced plans to integrate AI-assisted reviewing, using LLM-generated opinions as an additional perspective for expert evaluation[2].

Current LLM-based automatic reviewing methods generally fall into two main categories: (1) Prompt-based approaches, which leverage the in-context learning capabilities of LLMs to generate review content without parameter updates. These include direct prompting methods (e.g., AI-Scientist (Lu et al. 2024), which uses GPT-4 with tailored review templates) and multi-agent collaboration frameworks, where multiple LLM instances simulate interactions between reviewers and authors (e.g., AgentReview (Jin et al. 2024) and ReviewMT (Tan et al. 2024)). (2) Fine-tuned approaches, which train open-source LLMs (e.g.,

[*]These authors contributed equally.
[†]Corresponding authors.

[1]https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/
[2]https://aaai.org/conference/aaai/aaai-26/main-technical-track-call/

LLaMA (Touvron et al. 2023) or Qwen (Bai et al. 2023) series) on peer review data to align the model's outputs with expert reviewing criteria (e.g., CycleReviewer (Weng et al. 2025) and DeepReview (Zhu et al. 2025)).

Despite recent progress in automatic paper reviewing, existing methods still face several notable challenges. Prompt-based approaches, which rely solely on LLMs and instructions, often generate superficial and shallow review content (Zhou, Chen, and Yu 2024). Studies have shown that LLMs tend to produce generally positive but low-discriminative evaluations, failing to capture the nuanced reasoning required in peer review (Feng, Sun, and You 2025). Moreover, these methods are highly sensitive to prompt design (Errica et al. 2025); small changes in prompt wording can lead to significantly different outputs, resulting in limited stability and robustness. On the other hand, fine-tuned approaches that adapt open-source LLMs using peer review data suffer from data scarcity and bias. Publicly available high-quality review datasets are limited in size and scope, which restricts the generalizability of trained models. Additionally, these methods typically produce only a single-perspective review, lacking the ability to model multi-reviewer interactions and argumentative dynamics inherent in real-world peer review. Furthermore, both categories remain susceptible to hallucinations, wherein LLMs generate factually incorrect or misleading content, compromising fairness and reliability (Huang et al. 2025).

To address these challenges, we propose **ReViewGraph**, a novel framework for automatic paper reviewing that performs heterogeneous graph reasoning over LLM-simulated reviewer-author debates. Our approach explicitly captures multi-perspective opinions and their discourse-level interactions, thereby enhancing the depth, interpretability, and controllability of automated reviewing.

ReViewGraph begins by simulating multi-round reviewer-author debates through a multi-agent collaboration framework, resulting in a rich set of opinion exchanges. We then construct a heterogeneous debate graph to represent these interactions, comprising four node types: *Title*, *Evaluation Dimension*, *Reviewer Opinion*, and *Author Opinion*, and four meta-relation types that encode review structure and argumentative dynamics. These include (1) paper-to-dimension associations, (2) reviewer opinions tied to specific evaluation criteria (e.g., *Methodological Novelty*, *Motivation Clarity*, *Experimental Completeness*, *Writing Fluency*), (3) inter-reviewer relations such as *agree*, *disagree*, and *complement*, and (4) reviewer-author interactions like *clarify*, *reject*, and *accept*. To instantiate this graph, we use in-context prompting to extract opinion triplets and classify opinions into evaluation dimensions. This structured representation enables ReViewGraph to perform fine-grained relational reasoning and make informed final decisions. Building on this structured graph, we further apply a heterogeneous graph Transformer to perform relational reasoning and predict the final review decision.

To evaluate the effectiveness of ReViewGraph for automatic paper reviewing, we collect three benchmark datasets from OpenReview and compare our approach against seven strong baseline methods. Experimental results demonstrate that ReViewGraph consistently outperforms all baselines across these datasets. Notably, it achieves an average relative improvement of 15.73% over the second-best baselines. Overall, the main contributions of this work are: (1) We propose **ReViewGraph**, a novel framework for automatic paper reviewing that models reviewer–author interactions as heterogeneous graphs constructed from LLM-simulated multi-round debates. (2) We design a structured heterogeneous debate graph with semantically-typed nodes and edges to capture fine-grained argumentative relations across diverse review perspectives, and leverage a graph neural network to perform relational reasoning over this structure. (3) Extensive experiments on three datasets show that ReViewGraph consistently outperforms 7 strong baselines, achieving an average relative improvement of 15.73% over the second-best models.

Related Works

Traditional Automatic Reviewing. Early efforts in automatic paper reviewing primarily relied on manually curated review data and traditional neural classifiers for acceptance prediction. Kang et al. (2018) introduced the PeerRead dataset and trained classifiers using manually labeled reviews. Ghosal et al. (2019) incorporated sentiment features to improve prediction accuracy. To move beyond simple acceptance prediction, researchers began exploring automatic review generation. Bartoli et al. (2016) proposed one of the earliest neural frameworks, trained on 48 papers from their lab, to generate review comments. Nagata (2019) generated sentence-level feedback on grammar errors to support academic writing. ReviewRobot (Wang et al. 2020) constructed a knowledge graph from the input paper, predicted review scores, and selected templated comments based on both scores and supporting evidence. However, traditional neural models struggled with long and technical documents, lacking the capacity for deep semantic understanding, which ultimately hindered progress in this field.

LLM-based Automatic Reviewing. Recent advancements in LLMs have inspired a growing body of work on LLM-based automatic reviewing. Multi-agent collaboration frameworks aim to simulate the multi-role dynamics of real-world peer review. For example, ReviewMT (Tan et al. 2024) and AIScientist (Lu et al. 2024) reframe the review process as a multi-round, long-context dialogue among multiple roles, i.e., reviewers, authors, and meta-reviewers. AgentReview (Jin et al. 2024) further explores this direction by prompting LLMs to assume diverse reviewer personalities and decision strategies. Fine-tuned reviewer models enhance alignment with human review standards through supervised training. CycleReviewer (Weng et al. 2025) fine-tunes an open-source LLM on domain-specific review data, simulating multiple reviewers who assess the paper across different dimensions. DeepReview (Zhu et al. 2025) extends this line of work by modeling multi-step reviewer reasoning and training on generated rationales to produce more coherent, logically grounded reviews. Graph-based approaches such as GraphEval (Feng, Sun, and You 2025) leverage LLM prompting to segment abstracts into discrete opinion sen-

tences, which are then connected via similarity-based edges and processed through graph reasoning to predict acceptance decisions.

While these methods represent meaningful progress, they either treat interactions implicitly or lack fine-grained modeling of argumentative structures. In contrast, our proposed method, **ReViewGraph**, explicitly models reviewer-author debates as heterogeneous graphs, captures multi-perspective viewpoints, and performs structured reasoning over interaction relations using graph neural networks.

The ReViewGraph Framework

As illustrated in Figure 1, given a paper D, ReViewGraph first employs multi-agent collaboration to simulate multi-round debates between reviewers and the author. Based on the generated debate content, it then constructs a heterogeneous interaction graph and performs structured relational reasoning to predict the final review decision s for the paper. We provide a detailed explanation of ReViewGraph in the following.

Multi-agent Reviewer-Author Debate Simulation

To obtain a fine-grained understanding of the target paper, we propose a multi-agent collaboration framework that simulates the dynamics of real-world reviewer–author interactions. The framework consists of four role-specific agents: three regular reviewer agents, one author agent, and a senior reviewer agent who serves as a meta-level coordinator.

The simulation proceeds in three stages: (1) Initial Review Stage. Each regular reviewer agent is instantiated using a multimodal LLM capable of processing both textual and visual content. Given a set of review criteria and the full content of the target paper D, they are encouraged to first recognize the paper's strengths and substantive contributions, providing positive feedback where appropriate. At the same time, they are expected to raise concerns, highlight potential weaknesses, and offer critical analysis, especially for sections that appear ambiguous or underspecified. This stage ensures a comprehensive and balanced evaluation of the paper's strengths and limitations. (2) Author Rebuttal Stage. The senior reviewer agent then prompts the author agent to generate a point-by-point response to the reviewers' feedback. The author is instructed to clarify any misunderstandings, answer specific technical questions, and defend the paper's contributions where challenged. This stage aims to simulate the rebuttal phase commonly found in peer review, where authors have the opportunity to address regular reviewers' concerns. (3) Re-evaluation Stage. After receiving the author's rebuttal, the senior reviewer agent reminds the three regular reviewer agents to re-express or refine their opinions in light of the new information. Regular reviewers are encouraged to reassess their initial judgments, revise their critiques, or reaffirm their positions based on the clarified understanding.

This multi-round dialogue process enables the system to capture nuanced argumentative structures and simulate realistic peer review behaviors, laying the foundation for downstream review decisions.

Heterogeneous Debate Graph Construction

To represent the complex reasoning and interaction dynamics of the peer review process, we construct a heterogeneous debate graph from simulated reviewer-author interactions. Following the standard formulation of heterogeneous information networks (Hu et al. 2020), we define the graph as $G = \{\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}\}$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, $\mathcal{A}$ is the set of node types, and $\mathcal{R}$ is the set of edge types. Each node $v \in \mathcal{V}$ and $e \in \mathcal{E}$ is associated with a type via node and edge type mapping functions: $\psi(v) : V \rightarrow \mathcal{A}$, $\eta(e) : E \rightarrow \mathcal{R}$. Each edge corresponds to a **meta-relation** in the form $< \psi(s), \eta(e), \psi(t) >$, where s and t are source and target nodes.

Nodes. $A = \{$*Title*, *EvaluationDimension*, *ReviewerOpinion*, *AuthorOpinion*$\}$, instantiated as follows: (1) Title Node ($\psi(v) =$ *Title*): It represents the paper being reviewed, with its content set to full paper title (e.g., *"Automatic Paper Reviewing with Heterogeneous Graph Reasoning over LLM-Simulated Reviewer-Author Debates."*); (2) Evaluation Dimension Nodes ($\psi(v) =$ *EvaluationDimension*): They represent critical dimensions of academic review. To capture a holistic evaluation of a paper's quality, we focus on four core dimensions that are frequently emphasized in real-world peer reviews. These aspects reflect both the intellectual merit and presentational quality of a submission. We define four dimensions in this work: *Methodological Novelty*, *Experimental Completeness*, *Motivation Clarity*, *Writing Fluency*; (3) Reviewer Opinion Nodes ($\psi(v) =$ *ReviewerOpinion*): Each reviewer agent produces a set of comments, where each opinion in their comments is represented as a separate node. (4) Author Opinion Nodes ($\psi(v) =$ *AuthorOpinion*): The author agent responds to the reviewers with their opinions, which are also modeled as individual nodes.

Meta-Relations. We define four types of edges (i.e., relations) $\mathcal{R}$, each corresponding to a meaningful discourse connection. The resulting meta-relations $< \psi(s), \eta(e), \psi(t) >$ include: (1) Paper–Dimension Relations: $<$ *Title*, *has_aspect*, *EvaluationDimension* $>$, it represents that each paper is reviewed under several standard dimensions. (2) Dimension–Opinion Relations: $<$ *ReviewerOpinion*, *reviewed_by*, *EvaluationDimension* $>$. It means that each review opinion belongs to a specific evaluation dimension. (3) Inter-Reviewer Argumentative Relations: $<$ *ReviewerOpinion*, r, *ReviewerOpinion* $>, r \in \{agree, disagree, complement, progressive, independent\}$. These relations reflect viewpoint interactions among reviewers over shared dimensions. (4) Reviewer–Author Interaction Relations: $<$ *ReviewerOpinion*, r, *AuthorOpinion* $>, r \in \{accept, reject, clarify, compromise, extend, neutral\}$. These edges represent how authors respond to reviewer critiques.

Graph Instantiation. To instantiate the graph with the aforementioned meta-relations, we need to extract the relationships among reviewers' opinions as well as those between reviewers' and authors' opinions from their interactions. Additionally, it is necessary to categorize the evaluation dimensions to which each reviewer opinion belongs. To

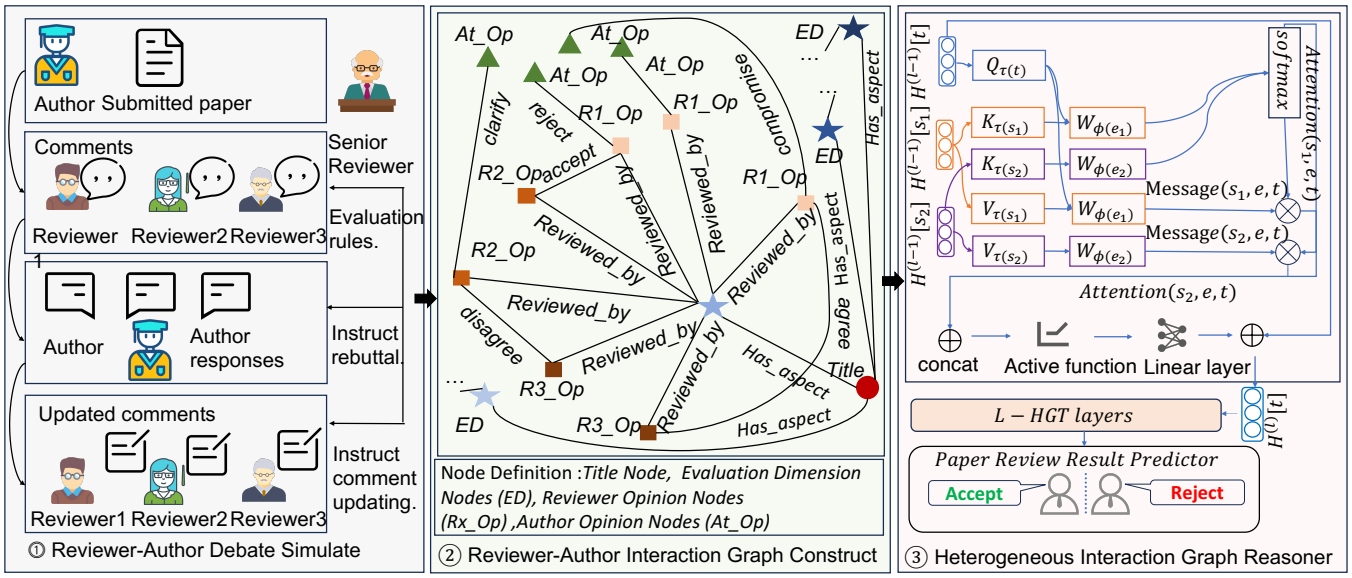


Figure 1: The workflow of ReViewGraph. Reviewer1, Reviewer2, and Reviewer3 represent three regular reviewer agents. Nodes are typed by shape: circles denote paper topic nodes, pentagrams denote evaluation dimension nodes (e.g., Methodological Novelty), squares denote opinions of the reviewers, and triangles represent the author's opinions.

achieve this, we first present the simulated dialogue between reviewers and the author, then prompt it to identify pairs of opinion statements and the relation that holds between them. For instance, if Reviewer A challenges a point raised by Reviewer B, the extracted triplet would reflect this disagreement relation between their respective opinions. Similarly, when an author responds to a reviewer's critique, the extracted triplet captures the response relation between the author's and the reviewer's views. Next, we further classify each opinion in the triplets into its corresponding Evaluation Dimension (e.g., *Methodological Novelty*, *Experimental Completeness*, *Motivation Clarity*, *Writing Fluency*). This is also achieved via in-context prompting, where the LLM is guided to assign a category to each opinion statement based on its content and argumentative context.

Through this pipeline, we construct a final heterogeneous debate graph that captures multi-perspective opinions and their semantic relationships within reviewer–author debates.

Reviewer-Author Debate Graph Reasoning

Given the reviewer-author debate graph constructed in the previous section, we adopt a Heterogeneous Graph Transformer (HGT) to perform reasoning over the structured interactions between reviewers and authors. Then, we can obtain contextualized representations of all nodes in the constructed heterogeneous graph, which can then be used for downstream tasks such as review result prediction.

Formally, for a target node $t \in V$, HGT learns its representation by performing heterogeneous mutual attention, heterogeneous message passing, and target-specific aggregation over each source node s connected to t through a relation type (edge). HGT usually consists of L stacked layers, which means the output of the l-th HGT layer, denoted as $H^{(l)}$, is fed as input into the next layer. Then the final nodes are represented as $H^{(L)}$.

Heterogeneous Mutual Attention. The Heterogeneous Mutual Attention mechanism determines the importance (i.e., weights) of neighboring nodes. For a target node t, its neighboring nodes $s \in N(t)$ may be connected via different meta-relations, represented as $< \psi(s), \eta(e), \psi(t) >$. Similar to the vanilla Transformer architecture, HGT maps the target node t to a query vector and each source node s to a key vector, using their dot product to compute attention scores. To model the diverse distributions of meta-relations, HGT further adopts relation-specific projection weights, allowing it to distinguish and effectively capture the semantics of different types of meta-relations. Specifically, in the attention mechanism of HGT, the weight matrices are decomposed into three components corresponding to the projections of the source node s, the edge e, and the target node t. Formally, for each neighboring node s of t, the attention mechanism computes multi-head attention scores, concatenates them, and applies the softmax function for normalization:

$$\mathbf{ATN}(s,e,t) = \sigma_{\forall s \in \mathcal{N}(t)}[\text{concat}_{i\in[1,Z]}(\text{attn}^{\text{h}_i}(s,e,t)] \quad (1)$$

$$\text{attn}^{\text{h}_i} = (K^i(s)W^{attn}_{\eta(e)}Q^i(t)) \cdot \frac{\mu_{<\psi(s),\eta(e),\psi(t)>}}{\sqrt{d}} \quad (2)$$

where Z is the number of attention head, and σ denotes the Softmax function. For the i-th attention head h_i, the key of the source node and the query of the target node are first interacted via a relation-specific projection matrix $W_{\eta(e)}$, scaled by a prior weight $\mu_{<\psi(s),\eta(e),\psi(t)>}$ representing importance for different meta relation triplets, and then divided by $\sqrt{d}$, following the standard Transformer scaling. Specifically, $\psi(s)$-type source node s is projected into

$K^i(s) \in \mathbb{R}^{d_h}$, and the $\psi(t)$-type target node t is projected into $Q^i(t) \in \mathbb{R}^{d_h}$, where $d_h = d/h$ denotes the dimensionality of each attention head. $W^{attn}_{\eta(e)}$ denotes the edge-based matrix for each edge type $\eta(e)$.

Heterogeneous Message Passing. This mechanism generates the information (i.e., messages) transmitted from neighboring nodes. Specifically, for (s, e, t), the multi-head message can be calculated as:

$$\mathbf{MSG}(s, e, t) = \text{concat}_{i\in[1,Z]} mesg^{h_i}(s, e, t) \quad (3)$$

$$mesg^{h_i}(s, e, t) = \text{Linear}^i_{\psi(s)}(H^{(l-1)}[s])W^{mesg}_{\eta(e)} \quad (4)$$

where $\psi(s)$-type source node s is projected into i-th message vector with the linear layer $\text{Linear}^i_{\psi(s)}$, and $W^{mesg}_{\eta(e)}$ is used to incorporate the edge dependency. Finally, all messages from Z heads are concatenated as the final message representation for each source node.

Target-Specific Aggregation. Once the attention weights and the messages from the source nodes of the target node t are computed, the representation of t at the l-th layer is updated as follows:

$$H^{(l)}[t] = \text{Linear}_{\psi(t)}(\lambda \hat{H}^{(l)}[t]) + H^{(l-1)}[t] \quad (5)$$

$$\hat{H}^{(l)}[t] = \oplus_{\forall s\in\mathcal{N}(t)} \mathbf{ATN}(s, e, t) \cdot \mathbf{MSG}(s, e, t) \quad (6)$$

where $\oplus$ denotes the weighted summation operation over all source nodes $\forall s \in \mathcal{N}(t)$ of the target node t, λ serves as a rescaling factor.

Review Result Prediction with HGT After obtaining the vector representation of each node in the heterogeneous debate graph between reviewers and authors, we first apply mean pooling over node embeddings grouped by their types. Since $\psi : V \to \mathcal{A}$ is a mapping from nodes $v \in V$ to their corresponding node types $a \in \mathcal{A}$, where $\mathcal{A}$ denotes the set of all possible node types. Then, the pooled representation for each node type $a \in \mathcal{A}$ is computed as:

$$h_a = \frac{1}{|V_a|} \sum_{v\in V, \psi(v)=a} H^{(l)}[t] \quad (7)$$

where $V_a = \{v \in V \mid \psi(v) = a\}$ denotes the set of nodes of type a. Next, we concatenate the pooled representations of all node types:

$$h_{\text{concat}} = ||_{a\in\mathcal{A}} h_a \quad (8)$$

where $||$ denotes vector concatenation. Finally, the concatenated vector is fed into a two-layer feedforward neural network to predict the final review decision (accept/reject):

$$\hat{y} = \text{Softmax}\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot h_{\text{concat}} + b_1\right) + b_2\right) \quad (9)$$

where W_1 and W_1 are learnable parameters, and b_1 and b_2 are biases.

Experiments

Experimental Settings

Task. Following prior works (Jin et al. 2024; Zhu et al. 2025; Weng et al. 2025; Feng, Sun, and You 2025), we use the full paper as input and aim to predict the final review decision. The decision s is selected from a predefined set of review outcomes: *accept* or *reject*.

Year	Mode	Accept (A-o / A-p / A-s)	Reject	Total
2023	Train	532 (42 / 349 / 141)	419	951
	Val	75 (6 / 49 / 20)	59	134
	Test	155 (13 / 101 / 41)	121	276
2024	Train	583 (42 / 361 / 180)	429	1012
	Val	82 (6 / 51 / 25)	61	143
	Test	169 (12 / 104 / 53)	123	292
2025	Train	712 (107 / 394 / 211)	631	1343
	Val	101 (15 / 56 / 30)	90	191
	Test	207 (32 / 114 / 61)	181	388

Table 1: Dataset statistics. Accept numbers are shown with their subcategory breakdown: Oral (A-o), Poster (A-p), Spotlight (A-s).

Datasets and Evaluation Metrics. To validate the effectiveness of ReViewGraph, we crawled submission data and review outcomes from ICLR 2023, 2024, and 2025 available on OpenReview. The statistics of the collected data are shown in Table 1. To thoroughly assess how well the evaluation methods align with human reviewers, we report each method's performance in terms of Accuracy, Macro Precision, Macro Recall, and Macro F1 score.

Baselines and Implementation Details. We compare our proposed framework against several representative baselines that leverage LLMs for automatic paper review: (1) ICL-based Method (Brown et al. 2020): Directly prompts a pre-trained LLM using in-context examples to predict the review decision based on the full paper content. (2) CoT-based Method (Wei et al. 2022): Encourages the LLM to perform chain-of-thought reasoning before generating the final review outcome. (3) AI-Scientist (Lu et al. 2024): Redefines the review process as a multi-turn, long-context dialogue among multiple roles. (4) CycleReviewer (Weng et al. 2025): Fine-tunes an open-source LLM using domain-specific peer review data. (5) DeepReview (Zhu et al. 2025): Extends CycleReviewer by modeling multi-step reviewer reasoning and training on generated rationales. (6) GraphEval (Feng, Sun, and You 2025): Segments paper abstracts into discrete opinion sentences using LLM prompting, connects them via similarity-based edges to construct a graph, and applies graph-based reasoning to predict the final review decision. The implementation details of ReViewGraph and the baselines are provided in the source code repository.

Experimental Results

Overall Performance. As shown in Table 2, our proposed ReViewGraph consistently outperforms all baseline methods on the ICLR 2023, 2024, and 2025 datasets across four key metrics. Notably, ReViewGraph achieves its strongest performance on the ICLR 2025 dataset, with all metrics exceeding 70 and an average relative improvement of 15.73% over competitive baselines. To assess statistical significance, we performed two-sample T-tests on the accuracy and F1 scores of ReViewGraph and the best-performing baseline, CycleReviewer-70B. The resulting p-values were 0.0192 and 0.0067, respectively, both below the 0.05 threshold, confirming that the improvements are statistically significant.

Datasets	ICLR 2023 Papers				ICLR 2024 Papers				ICLR 2025 Papers			
Method\Metric	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
ICL-based Method	58.33	68.75	52.66	42.44	61.30	70.69	54.39	46.65	56.70	71.74	53.66	43.04
CoT-based Method	57.97	65.04	52.33	42.25	61.64	71.35	54.80	47.38	57.47	75.06	54.45	44.24
AI-Scientist	59.06	70.64	53.49	44.03	60.48	67.32	53.69	45.65	58.51	**78.13**	55.53	49.95
GraphEval	50.00	47.08	47.51	46.35	48.97	44.23	45.30	43.75	46.91	44.25	45.38	43.13
CycleReviewer-8B	51.33	66.67	25.85	37.25	47.83	66.67	21.12	32.08	49.05	55.42	23.35	32.86
CycleReviewer-70B	61.23	**82.43**	39.35	53.28	57.73	71.30	45.56	55.60	63.05	72.22	50.24	59.26
DeepReview-14B-Std	61.23	73.08	49.03	58.69	59.93	69.70	54.44	61.13	63.77	64.29	47.06	54.34
ReViewGraph	**70.29**	69.85	**69.92**	**69.89**	**66.10**	65.48	**65.73**	**65.54**	**71.65**	72.04	**72.01**	**71.65**

Table 2: Performance Comparison between our ReViewGraph and Other Methods (Acc: Accuracy, P: Macro Precision, R: Macro Recall, F1: Macro F1 Score). The **bold** number indicates the best performance, while the underlined number represents the second-best.

ICLR 2023 Papers				
Method\Metric	**Acc**	**P**	**R**	**F1**
ReViewGraph	**70.29**	**69.85**	**69.92**	**69.89**
– w/o Title	67.03	67.98	68.02	67.03
– w/o Eval	69.20	69.63	69.86	69.17
– w/o RAR	69.57	69.37	69.64	69.37
– w/o IRR	68.12	67.99	68.26	67.95
– w/o Hetero	68.12	68.48	68.71	68.07
ICLR 2024 Papers				
ReViewGraph	**66.10**	65.48	65.73	**65.54**
– w/o Title	64.73	**65.78**	**65.99**	64.71
– w/o Eval	65.75	65.60	**65.99**	65.48
– w/o RAR	65.75	65.25	65.55	65.28
– w/o IRR	65.75	65.41	65.77	65.39
– w/o Hetero	65.75	65.04	65.21	65.10
ICLR 2025 Papers				
ReViewGraph	**71.65**	**72.04**	**72.01**	**71.65**
– w/o Title	66.49	69.30	67.59	66.02
– w/o Eval	68.30	69.28	68.90	68.24
– w/o RAR	70.36	70.71	70.70	70.36
– w/o IRR	69.59	70.19	70.04	69.57
– w/o Hetero	70.62	71.41	71.15	70.59

Table 3: Results of ablation study.

We further analyze the sources of ReViewGraph's performance gains in the following sections.

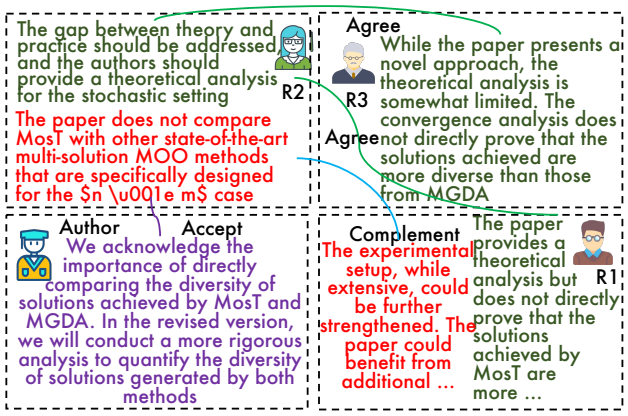
First, compared to prompt-based methods such as ICL-based and CoT-based methods, ReViewGraph demonstrates a substantial performance gain, especially in Accuracy and Macro-F1, with an average improvement exceeding 10 percentage points. This performance gap arises because prompt-based methods rely solely on paper content, limiting their ability to deeply understand and evaluate the review context. In contrast, ReViewGraph employs a multi-agent framework to simulate multi-turn reviewer–author debates, producing rich, structured, and multi-perspective opinion content.

Secondly, although AI-Scientist also models multi-role dialogues, ReViewGraph goes a step further by explicitly extracting semantic relations between viewpoints (e.g., acceptance, rejection, clarification, compromise). This enhances the semantic precision of interaction modeling, making the structural reasoning more logical and interpretable.

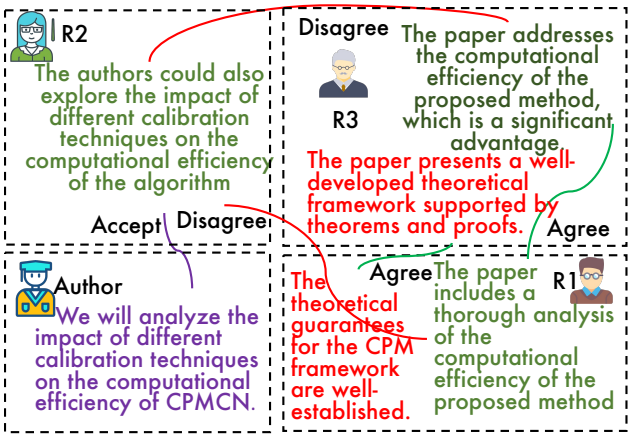
Third, ReViewGraph outperforms fine-tuned LLM-based baselines such as CycleReviewer and DeepReview. While CycleReviewer-70B demonstrates strong precision, its recall and overall consistency lag behind. For example, on the ICLR 2025 dataset, ReViewGraph achieves a Macro F1 score of 71.65, outperforming CycleReviewer-70B by more than 12 percentage points. Importantly, unlike these baselines that require resource-intensive fine-tuning, ReViewGraph operates without any LLM parameter updates. By leveraging a heterogeneous graph structure and applying Transformer-based reasoning, our method achieves greater generalizability, training efficiency, and inference controllability. Finally, compared to GraphEval, which constructs graphs based on sentence-level similarity within abstracts, ReViewGraph builds a more semantically expressive and structurally rich heterogeneous graph. It extracts explicit, labeled relations from full-text reviewer–author interactions, leading to higher representational fidelity and improved reasoning accuracy.

Ablation Analysis. We conducted ablation studies on three datasets to evaluate the impact of key components in our heterogeneous reviewer–author opinion graph. Specifically, we removed each of the following elements in turn: the paper title nodes (w/o Title), the evaluation dimension nodes (w/o Eval), the edges representing reviewer–author interactions (w/o RAR), and the edges capturing relations among reviewers' opinions (w/o IRR). Additionally, we replaced the heterogeneous graph structure with a homogeneous one by collapsing all node and edge types into a single type (w/o Hetero). In this homogeneous graph, all edges are uniformly labeled as *connected*, and nodes are not distinguished by type. This setup allows us to assess the benefits of explicitly modeling heterogeneity in opinion interactions.

As reported in Table 3, removing the paper title nodes (w/o Title) results in the most significant performance drop, indicating the importance of explicitly modeling the target submission. Excluding the evaluation dimension nodes (w/o Eval) also leads to a noticeable decline, confirming that capturing multiple review criteria enhances the model's discriminative capacity. Ablations that remove reviewer–author interaction edges (w/o RAR) or inter-reviewer relational edges (w/o IRR) both lead to moderate decreases, highlight-



(a) Case 1: Correct rejection prediction by ReViewGraph.



(b) Case 2: Correct acceptance prediction by ReViewGraph.

Figure 2: Two representative cases of ReViewGraph.

ing the value of modeling detailed argumentative and logical relations. Finally, replacing the heterogeneous graph with a homogeneous graph structure (w/o Hetero), where node and edge types are collapsed, further impairs performance, demonstrating the effectiveness of explicitly modeling heterogeneity in reviewer–author debates.

Case Study. To analyze the advantages of our method, we present two representative cases as follows: **Case 1: Correct Rejection Prediction Amid Subtle Negative Consensus.** In the first case shown in Figure 2a, the ground-truth decision was *Reject*. While the paper proposed a novel approach, multiple reviewers expressed concerns about the lack of rigorous theoretical analysis and inadequate comparisons with state-of-the-art baselines. Reviewer 1 noted the absence of evidence demonstrating the solution's diversity. Reviewer 2 emphasized the theoretical gap in stochastic settings and the lack of comparison with methods designed for the $n \gg m$ scenario. Reviewer 3 echoed similar concerns regarding limited convergence analysis and experimental comparisons. Although the authors acknowledged these issues and promised improvements, no concrete resolutions were provided. ReViewGraph successfully captured the negative consensus across multiple evaluation dimensions (e.g., theoretical rigor, baseline comparison) by modeling reviewer–reviewer agreement and reviewer–author interactions in a structured heterogeneous graph. The system correctly identified the dominant negative stance despite superficially polite language or minor positive comments, and inferred the rejection outcome. In contrast, methods such as ICL, CoT prompting, and LLM-based dialogue agents misclassified the case as *Accept*, likely due to over-reliance on isolated affirmative signals. **Case 2: Correct Acceptance Prediction Despite Isolated Criticism.** In another instance shown in Figure 2b, the ground-truth decision was *Accept*. Reviewer 2 suggested further exploration of the impact of calibration techniques on computational efficiency. However, both Reviewer 1 and Reviewer 3 explicitly stated that the paper adequately addressed computational concerns and praised its theoretical guarantees. Reviewer 2 ultimately agreed with the decision to accept. The authors also acknowledged the suggestion and expressed willingness to expand the analysis in future revisions. While several baseline models incorrectly predicted a *Reject* decision, possibly misinterpreting Reviewer 2's suggestion as a critical flaw, our model accurately inferred that Reviewer 2's comment was an isolated and non-decisive concern. By modeling the majority–minority stance across reviewers and the consistency between evaluation dimensions and final decisions, our method avoided overweighing minority dissent.

Conclusion

We proposed **ReViewGraph**, a novel framework for automatic paper reviewing that leveraged heterogeneous graphs constructed from simulated reviewer–author interactions. By modeling multi-agent debates, we extracted structured opinion relations among reviewers and authors and represented them as a heterogeneous graph with semantically typed nodes and edges. A heterogeneous graph Transformer was then applied to reason over this structure and predict final review outcomes. Through comprehensive experiments on three ICLR datasets, ReViewGraph consistently outperformed strong prompt-based, fine-tuned, and graph-based baselines across multiple evaluation metrics. Notably, our method achieved a relative improvement of 15.73% in Macro F1 score over the best-performing baseline, while requiring no LLM fine-tuning. Ablation studies further validated the importance of each graph component, particularly the modeling of evaluation dimensions and inter-agent relations. Case studies also demonstrated that ReViewGraph was able to correctly interpret nuanced reviewer dynamics, such as subtle consensus or isolated dissent, which were often misclassified by other methods. These findings highlighted that explicitly modeling reviewer–author interactions and discourse-level semantics led to more robust, interpretable, and context-aware automated reviewing. Overall, ReViewGraph offered a scalable and effective framework for augmenting peer review with structured LLM reasoning, pointing toward future directions in trustworthy AI-assisted scientific evaluation.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. Min Yang was supported by National Key Research and Development Program of China (2024YFF0908201), National Natural Science Foundation of China (Grant No. 62376262), the Natural Science Foundation of Guangdong Province of China (2024A1515030166, 2025B1515020032). Xian Wei is supported by the General Program of Shanghai Natural Science Foundation (No.24ZR1419800, No.23ZR1419300), the National Natural Science Foundation of China (No.42130112, No.42371479), the Science and Technology Commission of Shanghai Municipality (No.22DZ2229004), and Shanghai Frontiers Science Center of Molecule Intelligent Syntheses. Shiwen Ni was supported by GuangDong Basic and Applied Basic Research Foundation (2023A1515110718 and 2024A1515012003), China Postdoctoral Science Foundation (2024M753398), Postdoctoral Fellowship Program of CPSF (GZC20232873) and Shenzhen Science and Technology Program (JCYJ20250604182917023).

References

Alberts, B.; Hanson, B.; and Kelner, K. L. 2008. Reviewing peer review.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. *CoRR*, abs/2309.16609.

Bartoli, A.; Lorenzo, A. D.; Medvet, E.; and Tarlao, F. 2016. Your Paper has been Accepted, Rejected, or Whatever: Automatic Generation of Scientific Paper Reviews. In *Availability, Reliability, and Security in Information Systems - IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2016, and Workshop on Privacy Aware Machine Learning for Health Data Science, PAML 2016, Salzburg, Austria, August 31 - September 2, 2016, Proceedings*, volume 9817 of *Lecture Notes in Computer Science*, 19–28. Springer.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Drozdz, J. A.; and Ladomery, M. R. 2024. The peer review process: past, present, and future. *British Journal of Biomedical Science*, 81: 12054.

Errica, F.; Sanvito, D.; Siracusano, G.; and Bifulco, R. 2025. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 1543–1558. Association for Computational Linguistics.

Feng, T.; Sun, Y.; and You, J. 2025. GraphEval: A Lightweight Graph-Based LLM Framework for Idea Evaluation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ghosal, T.; Verma, R.; Ekbal, A.; and Bhattacharyya, P. 2019. DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 1120–1130. Association for Computational Linguistics.

Hu, Z.; Dong, Y.; Wang, K.; and Sun, Y. 2020. Heterogeneous Graph Transformer. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 2704–2710. ACM / IW3C2.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2): 42:1–42:55.

Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; and Wang, J. 2024. AgentReview: Exploring Peer Review Dynamics with LLM Agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 1208–1226. Association for Computational Linguistics.

Kang, D.; Ammar, W.; Dalvi, B.; van Zuylen, M.; Kohlmeier, S.; Hovy, E. H.; and Schwartz, R. 2018. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 1647–1661. Association for Computational Linguistics.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J. N.; Clune, J.; and Ha, D. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *CoRR*, abs/2408.06292.

Nagata, R. 2019. Toward a Task of Feedback Comment Generation for Writing Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 3204–3213. Association for Computational Linguistics.

Stelmakh, I.; Shah, N. B.; Singh, A.; and III, H. D. 2021. Prior and Prejudice: The Novice Reviewers' Bias against

Resubmissions in Conference Peer Review. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1): 75:1–75:17.

Tan, C.; Lyu, D.; Li, S.; Gao, Z.; Wei, J.; Ma, S.; Liu, Z.; and Li, S. Z. 2024. Peer Review as A Multi-Turn and Long-Context Dialogue with Role-Based Interactions. *CoRR*, abs/2406.05688.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Wang, Q.; Zeng, Q.; Huang, L.; Knight, K.; Ji, H.; and Rajani, N. F. 2020. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, 384–397. Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Weng, Y.; Zhu, M.; Bao, G.; Zhang, H.; Wang, J.; Zhang, Y.; and Yang, L. 2025. CycleResearcher: Improving Automated Research via Automated Review. In *The Thirteenth International Conference on Learning Representations*.

Zhang, J.; Zhang, H.; Deng, Z.; and Roth, D. 2022. Investigating Fairness Disparities in Peer Review: A Language Model Enhanced Approach. *CoRR*, abs/2211.06398.

Zhou, R.; Chen, L.; and Yu, K. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 9340–9351. ELRA and ICCL.

Zhu, M.; Weng, Y.; Yang, L.; and Zhang, Y. 2025. DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process. arXiv:2503.08569.