Generative Reviewer Agents: Scalable Simulacra of Peer Review

Nicolas Bougie[1], **Narimasa Watanabe**[1]
{nicolas.bougie,narimasa.watanabe}@woven.toyota

[1]Woven by Toyota

Abstract

The peer review process is fundamental to scientific progress, determining which papers meet the quality standards for publication. Yet, the rapid growth of scholarly production and increasing specialization in knowledge areas strain traditional scientific feedback mechanisms. In light of this, we introduce Generative Agent Reviewers (GAR), leveraging LLM-empowered agents to simulate faithful peer reviewers. To enable generative reviewers, we design an architecture that extends a large language model with memory capabilities and equips agents with reviewer personas derived from historical data. Our experiments demonstrate that GAR performs comparably to human reviewers in providing detailed feedback and predicting paper outcomes. Beyond mere performance comparison, we conduct insightful experiments, such as evaluating the impact of reviewer expertise and examining fairness in reviews. By offering early expert-level feedback, typically restricted to a limited group of researchers, GAR democratizes access to transparent and in-depth evaluation.

1 Introduction

Assessing the quality of research is central to the advancement of scientific discovery. Peer review remains a cornerstone of scientific publication, ensuring that manuscripts meet standards of novelty, rigor, and significance. Although essential, this process faces several challenges, including biases (Stelmakh et al., 2021), inconsistencies among reviewers (Kravitz et al., 2010), and an urgent need for scalable solutions (Liu and Shah, 2023). Estimates suggest that researchers collectively invest millions of hours in reviewing activities annually (American Journal Experts (AJE), 2024). Furthermore, access to high-quality feedback remains limited to a small fraction of researchers with established networks. Large language models (LLMs) hold considerable potential in relieving some of these issues in the scientific review process.

Recent breakthroughs in LLMs have shown promise in human behavior modeling, enabling the creation of autonomous agents (Hardy et al., 2023; Jansen et al., 2023). A growing body of research has explored LLM-based agents for simulating diverse societal environments (Park et al., 2023; Törnberg et al., 2023), such as software engineering, and recommender system evaluation (Wu et al., 2023; Anonymous, 2024). However, studies examining the use of LLM-based agents for academic peer review remain sparse.

Few approaches have explored the use of LLMs as tools to assist researchers at various stages of the scientific workflow (Lu et al., 2024). Yet, the peer review process remains a particularly challenging domain (Jansen et al., 2023). For instance, ReviewerGPT has demonstrated how LLMs can identify errors, verify checklists, and select the best version of a paper (Liu and Shah, 2023). Other efforts have shown LLMs capable of reviewing academic manuscripts (Liang et al., 2024b) and generating creative research ideas (Koivisto and Grassini, 2023). Although LLM-generated reviews may be preferred by authors (Tyser et al., 2024), their ability to predict final paper outcomes still lags behind human experts (Lu et al., 2024). In addition, several challenges remain open. These include modeling the intricate relationships between ideas, claims, and technical details in lengthy and complex papers, accurately capturing granular reviewer profiles, and reliably predicting final acceptance outcomes. Addressing these issues is essential to achieving reviewers that match the nuance, diversity, and rigor of human judgment.

We present **G**enerative **A**gent **R**eviewers (GAR), a novel framework that simulates peer reviewers through LLM-based agents. Each agent is initialized using real-world datasets and equipped with four core modules: profile, memory, novelty, and

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 98–116

review modules. The profile module stores traits and historical preferences, including characteristics like *strictness* and *focus areas*, inferred from past reviews via *contrastive comparison*. The review process begins by constructing a graph-based representation of the manuscript, mapping relationships among ideas, claims, and results. Next, the novelty module assesses the manuscript's novelty with support from external knowledge. Leveraging this representation alongside retrieved genuine reviews from the memory module, the reviewer module then generates structured feedback and an overall score, which is conducted over multiple rounds. This process emulates real-world peer review where reviewers provide initial evaluations and refine them in subsequent iterations. Finally, a meta-reviewer agent synthesizes individual reviews to determine the paper's final decision.

2 Related Work

Advances in AI have introduced new methods that enhance the research process (Xu et al., 2021). Furthermore, (Wang et al., 2024a; Baek et al., 2024) LLMs can generate research concepts, and (Wang et al., 2024b) automate survey writing. LLMs have also extended to peer review tasks (Zheng et al., 2023; Wang et al., 2023; Miret and Krishnan, 2024), leveraging their language understanding to simulate reviewer decision-making (Liu et al., 2023). Studies show GPT-generated reviews align with human assessments (Robertson, 2023; Liang et al., 2023), while ReviewerMT (Tan et al., 2024) reformulates peer review as a multi-turn dialogue. Recent work includes simulating peer review through LLM agents (Jin et al., 2023), examining LLM reliability in review settings (Anonymous et al., 2023b), exploring LLMs as evaluators (Anonymous et al., 2023c), and showing LLMs can provide feedback but struggle with complex manuscripts (Anonymous et al., 2023a). Earlier approaches focused on semi-automated review tools (Checco et al., 2020), while recent systems like MARG (D'Arcy et al., 2024) use collaborative frameworks with distinct agents for different paper sections. GAR differentiates itself by incorporating a memory-based reviewer with granular personas and a graph-based paper representation that systematically connects evidence with arguments. By linking claims with evidence and retrieving relevant documents at a claim level, GAR enables memory-augmented multi-round evaluations, producing more comprehensive, faithful, and token-efficient peer review simulacra.

3 Methodology

Generative reviewers provide automated paper review, generating scores (soundness, presentation, contribution, overall, confidence), identifying strengths/weaknesses, and predicting acceptance.

Task Formulation. Given a paper $p \in \mathcal{P}$ and a reviewer $r \in \mathcal{R}$, let $y_{rp} = 1$ denote that reviewer r has reviewer the paper p, and subsequently assigned a score s_{rp} with $s_{rp} \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The average score of each paper p can be represented by $R_p = \frac{1}{\sum_{r \in \mathcal{R}} y_{rp}} s_{rp} \cdot y_{rp}$. The simulator's goal is to faithfully distill the human genuine preferences such as $\hat{y_{rp}}$ and $\hat{s_{rp}}$ of reviewer r for an unseen paper p.

Review Process Design. GAR employs a 4-phase pipeline to simulate the peer review process. **1. Graph Construction:** The manuscript is structured into a knowledge graph that establishes connections between essential ideas, claims, technical details, and results. **2. Reviewer Selection:** Next, three to six reviewers are selected and their profile modules are initialized from historical data. **3. Reviewer Evaluation:** Each manuscript undergoes a multi-round evaluation by independent reviewers. **4. Meta-Review:** Finally, a meta-reviewer compiles the reviews to determine the final decision.

3.1 Graph-Paper Representation

Parsing scientific manuscripts presents challenges due to their length and complex relationships between evidence and arguments. Contributions and technical details typically appear in early sections, while supporting results are often presented later, raising questions about information structuring, manuscript length management, and redundancy reduction. This raises several key questions: How can the diverse elements be effectively organized to enable LLM-based agents to cross-reference and analyze them? How can redundant claims or findings be minimized to ensure an accurate and thorough assessment? To escape these pitfalls, we introduce a graph-based representation $\mathcal{G}$ that organizes academic paper content into a structured graph $\mathcal{G}(p)$:

Acronym Extraction The first step identifies acronyms and their definitions from the manuscript. The LLM parses the title, abstract, and introduction to retrieve a list of acronyms and their correspond-

ing definitions, R_{acr}.

Extraction of Core Elements: The second step identifies and extracts instances of graph nodes and edges from each chunk of the source paper. Let $C^* = \{c_1^*, c_2^*, \ldots, c_n^*\}$ denote the set of chunks in the paper p (e.g., Introduction, Methods, Results). We leverage a multipart LLM prompt that first identifies all entities in the text, including *ideas*, *claims*, *technical details*, and *supporting evidences*, before identifying all relationships between clearly-related entities, including the source and target entities and a description of their relationship. Each entity $e \in E$ becomes a node in $\mathcal{G}(p)$, with relationships $r \in R$ as edges such as "proves" or "supports". The graph is defined as: $\mathcal{G}(p) = (E, R)$.

Concept Merging: To reduce redundancy, we merge nodes that represent the same or similar concepts but are phrased differently across the manuscript, using the function: $E' = LLM(\langle Q_{merge}, E, R_{acr} \rangle)$, where Q_{merge} is the merging prompt and (E') is the new entity set. If two claims are merged, their technical details and supporting evidence — edges, will then point to the newly merged entity, R'. The updated graph becomes: $\mathcal{G}'(p) = (E', R')$, and $\mathcal{G}(p) \leftarrow \mathcal{G}'(p)$.

Community Detection: Given the homogeneous undirected weighted graph $\mathcal{G}(p)$, we leverage a community detection algorithms to partition the graph into communities of nodes with stronger connections to one another than to the other nodes in the graph, with c referring to a community. Namely, we use Leiden (Traag et al., 2019) to partition the graph into modular communities of closely related nodes (Edge et al., 2024), grouping nodes into thematically related clusters.

Community-Based Descriptor: The final step creates report-like descriptors for each community in the Leiden hierarchy, $\hat{C} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k\}$. The representation of the nodes and edges in the community serves to query the LLM, which produces a descriptor $\hat{c}_i$ representing c_i. Each community in the graph is assigned its corresponding descriptor, $\hat{c}_i = LLM(\langle Q_{sum}, c_i, R_{acr} \rangle)$, where Q_{sum} is a prompt instructing the LLM to describe the community, its structure, and cite the original text as much as possible to mitigate hallucination. These descriptors are attached to the graph $\mathcal{G}(p)$.

3.2 Reviewer Agent Architecture

Leveraging this graph-based representation, GAR structures agents in terms of four specialized modules tailored for review scenarios.

3.2.1 Profile Module

The **profile module** ensures the alignment of synthetic agents with the diverse behaviors of genuine reviewers. Each reviewer persona has eight core attributes: **strictness**, **expertise level**, **focus areas**, **evidence focus**, **open-mindedness**, **ethic focus**, **tone**, and **attention to technical details**. Predicting these characteristics is inherently challenging, as anonymization in blind review limits each reviewer to a single evaluation. Thus, we introduce *contrastive comparison*, performing pairwise comparisons across inter-reviewer and intra-reviewer assessments. Specifically, we conduct N comparisons in which the LLM assesses whether the reviewer's review, $\bar{r}$, is stricter than another review from a different paper $\bar{r}_i$. To ensure fairness — acknowledging that stricter reviews are often associated with lower-quality submissions, the LLM is also presented with inter-reviews (anchors) of the same target paper as context. The strictness score s_r^* of the reviewer r is formally defined as follows:

$$s_r^* = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(\text{LLM}\left(r \succ r_i \,\middle|\, \langle Q_{comp}, \bar{r}, \bar{r}_{int}, \bar{r}_i \rangle\right)\right) \tag{1}$$

where Q_{comp} is the comparison prompt, $\bar{r}$ is the target reviewer's review, $\bar{r}_{int}$ represents the intra-reviews, and $\bar{r}_i$ refers to a review randomly sampled from another reviewer r_i. Finally, strictness is categorized into low, $medium$, and $high$ levels based on percentiles. Similar formulas apply for other characteristics. The **expertise level** is derived from real reviews using their confidence scores $\in \{1, 2, 3, 4, 5\}$ while **focus area** is extracted from past reviews via a one-shot prompt Q_{focus}.

3.2.2 Novelty Module

The **novelty module** draws upon external knowledge sources to gauge the originality of the manuscript in comparison to prior research. It begins by extracting keywords from the introduction of the targeted paper, which are then employed in a semantic search to retrieve similar papers (Ammar et al., 2018), $\mathcal{B}_{sim}$. Retrieved documents are filtered to include only prior work based on submission year. The *title*, *abstract*, and *introduction* of the retrieved papers serve to analyze the extent of innovation, clarity of differences from past contributions, and adequacy of related work citation. The LLM generates a novelty s_{nov} score from 1-4, accompanied by a concise explanation e_{nov}, formalized as: $(s_{nov}^*, e_{nov}^*) \leftarrow$

$LLM(\langle Q_{novel}, R_{acr}, \mathcal{B}_{sim}, p\rangle)$, where p denotes the source paper.

3.2.3 Memory Module

We present a novel **memory module** to enable retrieval-augmented reviews. Assuming a benchmark dataset, each academic paper is structured as a graph $\mathcal{G}(p) = (V, E)$, as detailed in Section 3.1. However, here, we introduce an extra step. Given a community descriptor $\hat{c} \in \hat{C}$, we query the LLM to determine if $\hat{c}$ is mentioned in the human review. If mentioned, the agent is instructed to cite the original review, otherwise, the LLM is prompted to output `No specific mention was found in the review.`, denoted as r_c. Then, the memory is filled with pairs of community descriptor $\hat{c}$ and associated reviews, $\{\hat{c}, r_c\}$. All descriptors $\hat{c}$ are embedded and used as index of the memory module, $\mathbf{h}_{\hat{c}}$. The memory offers two retrieval schemes, serving at different stages of our framework:

Community-level retrieval: Retrieve similar communities and their associated reviews $\{\hat{c}, r_c\}$, using this similarity function: $\text{sim}(\mathbf{h_c}, \mathbf{h}_{\hat{c}'}) = \frac{\mathbf{h}_{\hat{c}}^\top \mathbf{h}_{\hat{c}'}}{\|\mathbf{h}_{\hat{c}}\|\|\mathbf{h}_{\hat{c}'}\|}$, where $\hat{c}$ is the target community descriptor and $\hat{c}'$ represents other communities.

Paper-level retrieval: Retrieve similar papers based on node and edges overlap, comparing two papers p_1 and p_2 at the descriptor level rather than direct embedding similarity (Cohan et al., 2020). The similarity function $\text{sim}_{\text{struct}}$ between papers p_1 and p_2 is expressed as: $\text{sim}_{\text{struct}}(\mathcal{G}(p_1), \mathcal{G}(p_2)) = \frac{|\{\hat{c}\in\hat{C}_1 | \exists \hat{c}'\in\hat{C}_2 \mathbb{1}(\text{sim}(\mathbf{h}_{\hat{c}}, \mathbf{h}_{\hat{c}'})>\tau)\}|}{\max(|\hat{C}_1|, |\hat{C}_2|)}$, where τ is a scalar that defines whether communities discuss similar concepts, and $\hat{C}_1$ and $\hat{C}_2$ are community descriptors of p_1 and p_2.

3.2.4 Review Module

We enhance the agent's reasoning capabilities through Chain-of-Thought (Wei et al., 2022). Namely, the review is initiated by the agent processing the paper and generating an **initial review** $R_{r,0}$ based on its persona and the preliminary novelty assessment $\{s_{nov}, e_{nov}\}$. In this stage, the agent evaluates each *community descriptor* $\hat{C} \in \mathcal{G}(p)$, then outputs numerical scores (soundness, presentation, contribution, overall, confidence), weaknesses and strengths, as well as a preliminary binary decision. The initial assessment prompt is: $Q_{r,0} = \langle Q_{review}, Q_{novelty}, Q_{style}, s_{nov}, e_{nov}, R_{acr}, \hat{C}\rangle$, where the score, accompanied by a summary of supporting arguments, is formatted into plain text and passed on to subsequent review stages. Second, the agent r engages in **multi-round refinement**, where at turn k, it receives the review and thoughts from the previous response $R_{r,k-1}$. Agents are successively presented each community descriptor $\hat{c}$ from the manuscript to review along with the M most similar communities $\hat{c}'_1, ..., \hat{c}'_M$ retrieved from the memory module and their associated reviews $r_{\hat{c}'_1}, ..., r_{\hat{c}'_M}$. To reduce the cognitive workload of reviewers, the agent evaluates communities in blocks of size $\frac{|\hat{C}|}{K}$, where K is the number of review rounds. This retrieval-augmented scheme guides the agent in assessing each community by guiding its attention toward relevant human-like considerations. Drawing inspiration from retrieved exemplars, the agent may decide to add strengths, weaknesses or correct potential mistakes made during the initial review. Thus, the prompt at turn k is formally defined as:

$$Q_{r,k} = \langle Q_{check}, R_{r,0}, \bigcup_{k=1}^{K}(R_{r,k-1}, (\hat{c}_1, ..., \hat{c}_{\frac{|\hat{C}|}{K}}), (\hat{c}'_1, ..., \hat{c}'_M, \hat{c}'_2, ...), (r_{\hat{c}'_1}, ..., r_{\hat{c}'_M}, r_{\hat{c}'_2}, ...)\rangle, \tag{2}$$

where $R_{r,0}$ denotes the initial review and $R_{r,k-1}$ is the k-th review. Given the prompt $Q_{r,k}$, each review agent generates a response $R_{r,k}$, sampled from a probability distribution $R_{r,k} \sim P(\cdot|Q_{r,k})$, as well as the thoughts/rationales behind their choices. The last review is then selected as the final review of reviewer r of the paper p.

3.3 Meta-Reviewer

After the individual reviews are completed, a meta-reviewer agent synthesizes the final decision. It retrieves the top-K_2 most similar genuine papers and their meta-reviews, and combines them with the scores and reviews provided by the individual reviewers, along with its own preliminarily assessment. This produces a concise, structured summary highlighting the paper's strengths and weaknesses, consolidating key insights raised by individual reviewers and presenting a balanced evaluation of the submission. After T turns of self-reflection, the meta-reviewer generates the final decision R_{meta} following:

$$Q_{m,t} = \langle Q_{meta}, r_{\hat{K}_1}, ..., r_{\hat{K}_2} \bigcup_{t=1}^{T}(\overline{S_t}), \bigcup_{j=0}^{|\mathcal{R}|}(R_{j,K})\rangle \tag{3}$$

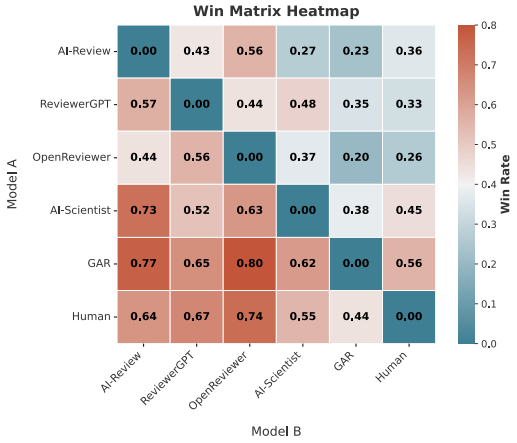

Figure 1: Win rates among LLM-generated and human reviewers based on GPT-4 preferences.

where $\hat{r_1}, ..., \hat{r_{K_2}}$ are retrieved meta-reviews, Q_{meta} is a meta-review prompt, and $\overline{S_t}$ is the meta-reviewer's summary of dialogues from turn t. The final acceptance decision is chosen among: `ACCEPT (ORAL)`, `ACCEPT (POSTER)`, or `REJECT`. We compare our meta-reviewer method, GAR, with a threshold-based approach $GAR^{>}$ that uses review scores against a fixed threshold reflecting conference acceptance ratios (see Sec. 4.3).

4 Experiments

Datasets. We primary conduct the experiments on the ICLR 2023 dataset, which consists of 3,797 papers obtained from Openreview, with additional experiments on ICLR 2022 and NeurIPS 2023 (Beygelzimer et al., 2021) datasets. Each paper was retrieved by at least three reviewers.

Baselines We compare our method with AI-Scientist (Lu et al., 2024), OpenReviewer (Tyser et al., 2024), ReviewerGPT (Liu and Shah, 2023), and AI-Review (Chiang and Lee, 2023).

Implementation. Each paper is evaluated by a committee of 3-6 reviewers and one meta-reviewer. Agents use GPT-4o-mini (OpenAI et al., 2024), with some experiments using GPT-4o (OpenAI et al., 2024) and Llama-3.1 (8b/70b) (Grattafiori et al., 2024). Results are averaged over 20 runs.

4.1 LLM vs Human Reviews

To assess the quality of LLM-generated reviews, five evaluators were given 200 papers, each with two anonymous reviews. As LLM Evaluators (Chiang and Lee, 2023) achieve comparable performance with human evaluators, we utilized GPT-4o as evaluator. For every paper, two reviewers were randomly assigned and evaluators were tasked with selecting their preferred review between the two provided for each paper. We ranked reviewers using a win matrix (Figure 1) and Bradley-Terry (BT) model coefficients. The win matrix records matchup outcomes, where element w_{ij} indicates the probability that competitor i defeats competitor j. Results in Table 1 show GAR leads with a score of 0.684, outperforming human reviewers (0.523). AI-Scientist and ReviewerGPT achieve scores of 0.242 and 0.000, respectively. Upon looking at the responses, GAR generated reviews stem from their depth, resulting in high preferences. One key factor is the retrieval of relevant reviews for each community descriptor, helping the LLM to identify specific issues or strengths. In contrast, some human reviews are more shallow, often due to reviewers having limited expertise in the field or constrained time for evaluating the manuscript.

Rank	Reviewer	Score
1	GAR	**0.684**
2	Human	<u>0.523</u>
3	AI-Scientist	0.242
4	ReviewerGPT	0.000
5	AI-Review	-0.365
6	OpenReviewer	-0.632

Table 1: Preference ranking of reviewers based on GPT-4o. **Bold**: best results; <u>underlined</u>: second-best.

Rank	Reviewer	Score
1	GAR	**0.143**
2	Human	<u>0.112</u>
3	AI-Scientist	0.000
4	AI-Review	-0.245
5	ReviewerGPT	-0.764
6	OpenReviewer	-1.461

Table 2: Human preference ranking of reviewers. The best results of each model are marked in **bold** and the second-best results are marked with <u>underline</u>.

4.2 Human Review Preferences

To evaluate GAR reviews against prior work, we followed the same experimental setup as Section 4.1, but with human evaluators selecting their preferred review between pairs, in one-on-one comparisons. Table 2 shows the ranking results. Similarly with LLM preferences, GAR achieves a top score of 0.143, higher than the human reviewers. On

Methods	NeurIPS		ICLR 22		ICLR 23	
	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑
Human*	0.66	0.49	0.66	0.49	0.66	0.49
Random Decision	0.50	0.33	0.50	0.38	0.50	0.40
Always Reject	0.50	0.00	0.50	0.00	0.50	0.00
AI-Scientist	0.58± 0.04	0.51± 0.06	0.65± 0.04	0.57± 0.05	0.63± 0.05	0.55± 0.06
OpenReviewer	0.39± 0.05	0.39± 0.04	0.49± 0.05	0.47± 0.05	0.50± 0.06	0.45 ± 0.05
ReviewerGPT	0.41±0.06	0.40±0.05	0.54±0.04	0.52±0.06	0.55±0.07	0.51±0.05
AI-Review	0.59±0.04	0.49±0.05	0.64±0.06	0.55±0.04	0.61±0.06	0.53±0.07
GAR	0.64±0.05	0.61±0.04	0.68±0.03	0.66±0.05	0.66±0.04	0.60±0.04
GAR$^{>}$	**0.68±0.05**	**0.62±0.05**	**0.71±0.04**	**0.67±0.06**	**0.70±0.05**	**0.69±0.05**

Table 3: GAR vs baselines on three datasets (1,000 papers each). **Bold**: best results; underline: second-best. GAR improvements are statistically significant ($p < 0.05$). (*) Human scores from (Beygelzimer et al., 2021). Results are averaged across 3 seeds.

the other hand, baseline methods underperformed (≤ 0) compared to human reviewers (0.112). A key advantage of GAR is the use of different reviewer personas, each focusing on distinct aspects of the paper. In contrast, most prior work struggle to determine specific criteria for assessment or to identify potential issues effectively at a claim level — they tend to produce generic feedback, as they do not explicitly retrieve relevant reviews.

4.3 Predicting Paper Acceptance

To evaluate the effectiveness of GAR, we compared its decisions against a ground truth dataset comprised of 1,000 papers from the NeurIPS 23, ICLR 22, and ICLR 23 submissions. The remaining reviews in each dataset were utilized to initialize the memory module. Table 3 shows our method outperforms previous state-of-the-art methods with an f1 score of 0.66 versus AI-Scientist's 0.54, significantly exceeding human reviewers' 0.49 score (Beygelzimer et al., 2021) ($p < 0.002$). We attribute GAR's improvements to our profile module with granular information (strictness, focus area) obtained via contrastive matching. Prior LLM-based methods struggle with complex papers and often overweight redundant claims. We alleviate this pitfall by leveraging the proposed paper representation, grouping together related claims/evidences and reducing redundancy through *concept merging*.

4.4 Human Likeliness

We use GPT-4o to assess whether agent-generated reviews appeared AI-generated or human-like using a 5-point Likert scale, with higher scores indicating stronger resemblance to human reviewers. As shown in Table 4, our method significantly outperforms AI-Scientist in generating reviews that align closely with human feedback. GAR scores are consistently higher, suggesting that the inclusion of graph-based memory and profile modules enhances the human-likeness of reviews. Additionally, allowing agents to simulate reviewer-specific characteristics, such as self-assessed confidence and depth of expertise, further contributed to review consistency and believability. Conversely, OpenReviewer and ReviewerGPT displayed tendencies towards inconsistency, such as generating shallow comments or narrowly focusing on methodological details without evaluating the validity of technical claims and results. This lack of critical depth raised suspicions of AI involvement.

	NeurIPS	ICLR 22	ICLR 23
AI-Scientist	3.34 ± 0.09	3.39 ± 0.11	3.38 ± 0.08
OpenReviewer	2.45 ± 0.10	2.43 ± 0.09	2.43 ± 0.09
ReviewerGPT	3.26 ± 0.13	3.25 ± 0.14	3.29 ± 0.15
AI-Review	3.30 ± 0.09	3.42 ± 0.11	3.38 ± 0.08
GAR	**3.89 ± 0.11***	**4.02 ± 0.10***	**3.99 ± 0.09***
Human	4.37 ± 0.08	4.45 ± 0.07	4.32 ± 0.08

Table 4: Human-likeness scores for several approaches.

4.5 Assistive Value for Human Reviewers

While the primary focus of our work has been on simulating peer review, an important question is whether GAR can assist human reviewers. To study this, we conduct a human-in-the-loop evaluation with 15 volunteer reviewers, each reviewing 6 ICLR-2024 papers. For half of the papers, reviewers had access to GAR-generated reviews, while the other half served as controls. After completing each review, participants answered three Likert-scale questions (1 = strongly disagree, 5 = strongly

Metric	Control	GAR-assisted	Δ	Direction
Review time (min) ↓	42 ± 11	34 ± 9	-8	Faster
Q1 Clarity ↑	3.2 ± 0.7	4.0 ± 0.6	$+0.8$	Higher
Q2 Confidence ↑	3.1 ± 0.8	3.7 ± 0.7	$+0.6$	Higher
Q3 Workload (rev.) ↑	2.9 ± 0.6	3.6 ± 0.5	$+0.7$	Lower Load

Table 5: Effect of GAR assistance on human reviewers. Likert items use a 5-point scale (1 = strongly disagree, 5 = strongly agree).

agree) on clarity of contributions, confidence in fairness, and perceived workload. We also recorded review time. As can be observed in Table 5, GAR reduces review time while improving clarity, confidence, and perceived workload. These findings reinforce the role of GAR as an assistive framework that augments, rather than replaces, human reviewers.

5 Conclusion

This research marks a step towards improving scientific writing and research by offering cost-effective, in-depth, and on-demand reviews. We describe an architecture for generative reviewers that employs a graph-based representation of manuscripts, a memory module for storing past reviews, and a novel technique to assign reviewers with specific traits and preferences. We then demonstrate the potential of GAR to achieve human-level feedback and accurately predict acceptance outcomes. Our vision is not to replace human reviewers but to enhance the review process by supporting them with synthetic reviewers capable of managing the increasing volume of submissions and providing early, constructive feedback. This collaboration between agents and human experts has the potential to accelerate scientific progress, foster innovation, and reduce time-to-publication.

6 Limitations

There are several limitations to our work. First, our study primarily focuses on isolating and evaluating specific factors in the peer review process, such as reviewer dedication or expertise, instead of accounting for the inherent variability and arbitrariness that occur in real peer review scenarios. Second, the large number of interacting modules makes it difficult to isolate the effect of each component; we include ablation studies in the Appendix to partially address this. In the future, we will explore and improve these aspects. Related to this, our analysis mainly isolates and examines individual variables of the peer review process. Real-world peer reviews, nevertheless, involve multiple interacting dimensions. Third, our method may inherit biases, due to the nature of LLMs. Our framework is largely reliant on the strengths and weaknesses of the underlying LLMs. The accuracy of the simulated user behavior may be impacted by LLMs' occasional inconsistent, biased, or unfounded outputs. A limitation is GAR's difficulty in evaluating highly novel or paradigm-shifting work, as noted in Appendix A.4. The system may struggle to recognize contributions that deviate from established norms, potentially overlooking groundbreaking ideas. Finally, our evaluation was conducted in the context of machine learning conferences; as such, the generalizability of our findings to other domains or conference communities remains to be established.

7 Ethics Statement

This paper introduces an LLM-powered framework that simulates peer reviewers to enhance the scientific review process. While GAR provides a scalable and efficient approach to assessing research quality and generating structured feedback, its deployment raises important ethical considerations.

First, LLM-based reviewer agents may inherit biases from their training data, potentially amplifying systemic biases in peer review, such as preferential treatment toward well-established institutions or underestimating novel but unconventional ideas. Additionally, these agents might struggle with subjective aspects of reviewing, such as evaluating the broader impact of a work, leading to inconsistencies in their assessments. Furthermore, reliance on automated reviewers raises concerns about the depersonalization of the review process, where human intuition, domain expertise, and contextual understanding remain irreplaceable.

Another ethical concern involves transparency and accountability. If LLM-generated reviews influence acceptance decisions, it is crucial to ensure that the decision-making process remains interpretable and that researchers understand the limitations of AI-generated feedback. Additionally, while GAR is designed to assist in peer review rather than replace human reviewers, there is a risk that institutions or conferences might over-rely on automated evaluations, reducing human oversight in critical decision-making processes. Automated reviewers should complement, rather than re-

place, human judgment, serving as tools to assist researchers rather than deterministic decision-makers. Recent machine learning conferences have reported an increase in reviews suspected to be AI-generated (Liang et al., 2024a). While LLM-generated reviews can provide valuable feedback, we strongly advise against their use as replacements for genuine reviewers in real-world peer review processes. Since LLMs are still prone to errors, human judgment remains essential to ensure the quality, fairness, and integrity of manuscript evaluations. By adhering to these principles, we aim to ensure that GAR contributes to a more efficient, fair, and reproducible peer review system.

References

American Journal Experts (AJE). 2024. The peer review process: 15 million hours of lost time. Accessed: 2024-10-25.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, and 4 others. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91. Association for Computational Linguistics.

Anonymous. 2024. SimUSER: When language models pretend to be believable users in recommender systems. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Anonymous and 1 others. 2023a. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Anonymous and 1 others. 2023b. Is llm a reliable reviewer? a comprehensive evaluation of llm on review-revision multiple-choice questions. *ACL Anthology 2024.lrec-main.816*.

Anonymous and 1 others. 2023c. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. The neurips 2021 consistency experiment.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Alessandro Checco and 1 others. 2020. Ai-assisted peer review. *Nature Communications*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

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

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Samir Haffar, Fateh Bazerbachi, and M. Hassan Murad. 2019. Peer review bias: A critical review. *Mayo Clinic Proceedings*, 94(4):670–676.

Mathew Hardy, Ilia Sucholutsky, Bill Thompson, and Tom Griffiths. 2023. Large language models meet cognitive science: Llms as tools, models, and participants. In *Proceedings of the annual meeting of the cognitive science society*, volume 45.

Bernard J Jansen, Soon-gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.

Yiqiao Jin and 1 others. 2023. Agentreview: Exploring academic peer review with llm agents. *arXiv preprint arXiv:2406.12708*.

Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601.

Richard L Kravitz, Peter Franks, Mitchell D Feldman, Martha Gerrity, Cindy Byrne, and William M Tierney. 2010. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PloS one*, 5(4):e10072.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, and 1 others. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *Preprint*, arXiv:2310.01783.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, and 1 others. 2024b. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. 2021. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8635–8643.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multimodal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

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

Santiago Miret and NM Krishnan. 2024. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. *arXiv preprint arXiv:2210.00105*.

Sotaro Shibayama, Deyun Yin, and Kuniko Matsumoto. 2021. Measuring novelty in science with word embedding. *PloS one*, 16(7):e0254034.

Jessie J Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Real ml: Recognizing, exploring, and articulating limitations of machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 587–597.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17.

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

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, and 1 others. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259*.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, and 1 others. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4).

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*.

A Appendix

A.1 Additional Details

In the profile module, the following attributes are derived from historical datasets:

- **Strictness** reflects the degree to which a reviewer adheres to high standards in evaluating submissions, ranging from lenient to highly critical.
- **Evidence Focus** describes how much importance the reviewer places on the evidence provided in the submission to support claims, highlighting their emphasis on empirical validation or theoretical soundness.
- **Open-mindedness** measures the reviewer's willingness to consider unconventional or novel ideas. A higher score indicates more openness to creative methodologies or speculative hypotheses.
- **Tone** refers to the overall style and approach taken by the reviewer in their feedback, ranging from highly critical to constructive.
- **Technical Focus** reflects the extent to which the reviewer is detail-oriented in evaluating the technical correctness and methodological rigor of a submission.

To enhance reliability of the novelty detection module, the LLM undergoes an *adversarial self-check* by re-evaluating its initial assessment in light of the retrieved documents and its prior reasoning. Namely, it is presented with the *same* context, including the retrieved documents and its own previously generated assessment, but is now instructed to identify any contradictory or overlapping evidence in prior work that might reduce or challenge the initial novelty determination. This iterative refinement results in a revised score, $s_{nov} \in \{1, 2, 3, 4\}$, and explanation e_{nov}, which are subsequently used during paper review to condition reviews. Unless otherwise specified in an experiment, all modules in GAR used `GPT-4o-mini`.

The meta-reviewer agent follows the same architecture as the reviewer agent, including the memory module and novelty module. The primary distinction lies in the initialization of its memory: whereas reviewer agents incorporate historical data from individual reviewers, the meta-reviewer agent's memory is initialized from historical meta-reviews. This enables the meta-reviewer agent to contextualize and emulate higher-level assessment consistent with human meta-reviewers.

Note that we filter out from the memory module any documents whose title matches the paper currently being reviewed, in order to avoid reusing prior reviews or content that could introduce evaluation bias. However, this approach is not foolproof: if a manuscript has undergone a title change or minor textual modifications, exact matching may fail to detect overlaps. Thus, contamination risks may remain, particularly for widely circulated preprints. Addressing this limitation may require more advanced strategies, such as fuzzy matching or content-based similarity filtering, which we leave to future work.

An example prompt when presenting claims, along with retrieved evidences and genuine feedback from the memory module is provided below:

Prompt Block

Idea 1: ($\hat{c}_1$) Utilizing graph neural networks (GNNs) to model user-item interactions in large-scale recommender systems. The approach claims to enhance scalability and accuracy through advanced message-passing mechanisms. Experiments indicate a 15% improvement in nDCG@10 compared to baseline collaborative filtering models on the MovieLens dataset.

Most Similar Claims:

- **Similar Claim 1:** ($\hat{c}_1'$) Implementing dynamic user-item graph construction for scalable recommendations using GNNs...
  Reviewer comment: ($r_{\hat{c}_1'}$) The dynamic graph approach is compelling but could benefit from further comparisons with static graph baselines.
- **Similar Claim 2:** ($\hat{c}_2'$) Applying attention-based GNNs to enhance explainability in recommender systems...
  Reviewer comment: ($r_{\hat{c}_2'}$) The work convincingly demonstrates improved explainability, but additional benchmarks against non-attention models are needed.

- **Similar Claim 3:** ($\hat{c}_3'$) Integrating GNNs with latent factor models to address cold-start issues in recommendation scenarios...
  Reviewer comment: ($r_{\hat{c}_3'}$) The integration with latent factors is innovative, though evaluations on datasets with extreme sparsity could strengthen the claim.

Idea 2: ($\hat{c}_2$) ...
Most Similar Claims: ...

where *Idea 1* refers to a descriptor of the paper being evaluated, *Similar claim 1* is a descriptor from another paper, which is similar to *Idea 1*, and *Reviewer Comment* is the actual review corresponding to this similar claim *Similar claim 1*. This structure helps agent reviewers to thoroughly assess each claim by guiding its attention toward relevant human-like considerations.

A.2 Simulation Platform

A challenging problem in processing research papers lies in preserving their structural integrity, particularly complex elements like mathematical formulas. Unlike prior work such as AI-Scientist (Lu et al., 2024), which converts PDFs to plain text and may compromise document formatting, we utilize Nougat (Blecher et al., 2023) to extract the Markdown (MMD) version of each manuscript, maintaining structural and formatting fidelity.

We argue that experimental results are indispensable for determining a paper's alignment with publication standards. Hence, we further graft figures into the paper representation. That is, figures that contain empirical findings, such as bar charts, are identified using Molmo-7b (Deitke et al., 2024). Next, we prompt GPT-4o to generate detailed captions for these figures, describing the methods being compared, important findings, and key results. Each caption is placed immediately following the original figure title, providing LLM reviewers with direct access to experimental data, enhancing their ability to rigorously evaluate the paper's technical soundness.

A.3 Bradley-Terry Model Details

In experiments 4.1 and 4.2, we measure and rank reviewers based on match outcomes, using a win matrix, coefficients from the Bradley-Terry (BT) model, and logistic regression. The win matrix records the results of matchups between competitors. For N competitors, the matrix W is an $N \times N$ grid where each element w_{ij} indicates the probability that competitor i defeats competitor j, calculated as $w_{ij} = \frac{\text{\# wins by } i \text{ over } j}{\text{total matches between } i \text{ and } j}$. The matrix is constructed by processing a list of match results, updating both win counts and total match counts for each competitor pair. The win matrix generated in our experiment is displayed in Figure 1.

The Bradley-Terry model applies a parametric approach to estimate the relative strengths of competitors through pairwise comparisons. In this model, the probability P that competitor m prevails over competitor m' is given by a logistic function: $P\left(H = \frac{1}{1+e^{\xi_{m'} - \xi_m}}\right)$, where ξ represents the vector of BT coefficients, with the constraint $\xi_1 = 0$ imposed. These coefficients are derived by minimizing the binary cross-entropy loss over all observed matches, using the following loss function: $\ell(h, p) = -\big(h \log(p) + (1-h)\log(1-p)\big)$. The optimization task can then be expressed as $\hat{\xi} = \text{argmin}_\xi \sum_{t=1}^{T} \ell\left(H_t, \frac{1}{1+e^{\xi_{A_2} - \xi_{A_1}}}\right)$, while keeping $\xi_1 = 0$ to anchor the scale. Once calculated, the BT coefficients ξ are used to rank competitors, ordering them from strongest to weakest by sorting the ξ values in descending order, as shown in Ranked Competitors = $\text{sort}\big(\xi, \text{descending}\big)$.

We developed an automated paper review framework, which demonstrates that LLMs, while still evolving, can provide review quality close to human standards. GPT-4o consistently produced the best results, occasionally reaching scores above human experts. However, we do not rely solely on proprietary models; as LLMs advance, both open and closed models are likely to improve. Our approach, therefore, remains model-agnostic, balancing the high performance of closed models like GPT-4o with the flexibility, lower cost, and transparency of open models such as Llama-3. Although open models currently show slightly lower quality, they hold the potential for cost-effective and adaptable AI systems. Future efforts will explore a closed-loop, self-improving system using open models to maximize discovery potential.

Another limitation concerns potential information leakage from pre-training corpora. Since most frontier LLMs are trained on proprietary datasets, it remains difficult to ascertain whether evaluation papers may have been partially included in the mod-

Methods	NeurIPS		ICLR 22		ICLR 23	
	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑
GAR (♦)	0.64±0.05	0.61±0.04	0.68±0.03	0.66±0.05	0.66±0.04	0.60±0.04
GAR$^{>}$ (♦)	**0.68±0.05**	**0.62±0.05**	**0.71±0.04**	**0.67±0.06**	**0.70±0.05**	**0.69±0.05**
GAR (♠)	0.60±0.06	0.58±0.05	0.65±0.05	0.63±0.04	0.64±0.05	0.57±0.06
GAR$^{>}$ (♠)	0.63±0.05	0.60±0.06	0.67±0.04	0.63±0.05	0.63±0.05	0.57±0.05
GAR (♥)	0.55±0.05	0.52±0.04	0.63±0.05	0.56±0.06	0.60±0.05	0.53±0.05
GAR$^{>}$ (♥)	0.58±0.05	0.53±0.04	0.65±0.06	0.59±0.04	0.61±0.06	0.64±0.06

Table 6: Effect of paper representation on the performance for predicting the acceptance of manuscripts. Asterisks (*) denote statistically significant improvements over the best baseline (t-test at $p < 0.05$).

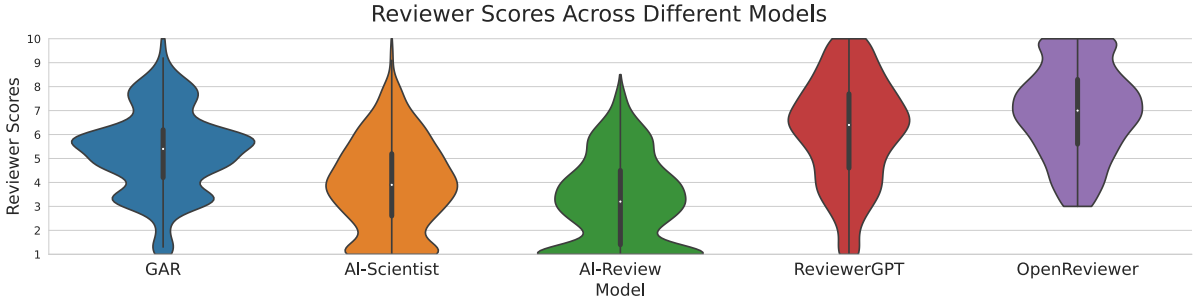

Figure 2: Violin plots showing the distribution of scores generated by several AI-generated models for ICLR 23 papers. Scores on the y-axis refer to paper ratings, which range from 1 (Strong Reject) to 10 (Weak Accept).

els' training data. Wwe acknowledge it as an inherent limitation of our approach. To mitigate this risk, future work could employ locally trained models on carefully controlled corpora, ensuring that training data is disjoint from evaluation papers.

A.4 Additional Experiments

A.4.1 Effect of Paper Representation

To validate the significance of our graph-based representation, we compare RAG's effectiveness in predicting paper acceptance using three input types: (1) community descriptors ♦, (2) node-edge representations ♠, and (3) raw text ♥. The experiment results, illustrated in Table 6, demonstrate that different manuscript representation methods significantly influence the F1-scores achieved in the review process. Our analysis shows that structured representations, particularly graphs and community-based descriptors, yield higher F1-scores compared to raw text. These structured formats enhance the extraction and alignment of key concepts, leading to more precise and contextually aware assessments. In contrast, raw text lacks such structured organization, resulting in lower F1-scores due to its reliance on linear narrative, which may increase the cognitive load for reviewers. This observation aligns with existing research in natural language processing, which underscores the value of structured representations in enhancing information extraction and interpretation (Liu et al., 2021; Schneider et al., 2022).

A.4.2 Review Scores Across Different Models

This experiment validates the effectiveness of five LLM-powered reviewers in scoring ICLR 23 papers: GAR, AI-Scientist, AI-Review, ReviewerGPT, and OpenReviewer. Figure 2 shows violin plots of score distributions aligned with ICLR ratings, allowing for an assessment of each model's alignment with human review standards. AI-Scientist exhibits a concentrated distribution around 4, indicating moderate alignment, while GAR showed greater variability. ReviewerGPT and OpenReviewer skewed higher, suggesting a higher inclination to accept manuscripts. On the other hand, GAR demonstrates consistent scores between 1 and 10, with the ability to accept high-quality papers (score >7) while strongly rejecting papers (≤3) that do not meet the quality standards for publication.

Methods	NeurIPS		ICLR 22		ICLR 23	
	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑
GAR (random persona)	0.59±0.05	0.57±0.05	0.60±0.07	0.68±0.06	0.61±0.03	0.61±0.06
GAR$^>$ (random persona)	0.62±0.05	0.59±0.04	0.63±0.04	0.69±0.05	0.64±0.06	0.65±0.06
GAR (historical persona)	0.65±0.05	0.60±0.04	0.65±0.06	0.64±0.04	0.69±0.06	0.66±0.05
GAR$^>$ (historical persona)	0.68±0.05	0.62±0.05	0.71±0.04	0.67±0.06	0.70±0.05	0.69±0.05
GAR (NN selected)	0.66±0.07	0.61±0.04	0.72±0.05	0.66±0.05	0.72±0.06	0.67±0.05
GAR$^>$ (NN selected)	**0.70±0.06**	**0.63±0.05**	**0.74±0.05**	**0.69±0.04**	**0.74±0.05**	**0.70±0.06**
GAR (w/o memory)	0.54±0.04	0.54±0.06	0.61±0.05	0.53±0.04	0.61±0.05	0.52±0.06
GAR$^>$ (w/o memory)	0.59±0.04	0.54±0.05	0.65±0.04	0.58±0.06	0.64±0.05	0.54±0.05
GAR	0.64±0.05	0.61±0.04	0.68±0.03	0.66±0.05	0.66±0.04	0.60±0.04
GAR$^>$	0.68±0.05	0.62±0.05	0.71±0.04	0.67±0.06	0.70±0.05	0.69±0.05

Table 7: Ablation study of GAR on three datasets, each consisting of 1,000 papers. Lines 3-5 report results for different approaches to select reviewers' persona. Line 6 highlights the performance without memory module.

A.4.3 Experiment: Reviewer Persona Selection

Reviewer personas shape the tone and content of paper evaluations. We compare three approaches for selecting reviewer personas:

- **Random Persona Selection**: Reviewer profiles are randomly selected among possible personas. This approach serves as the baseline, representing the scenario where the persona is not known in advance.
- **Historical Data Initialization**: Reviewer personas are initialized using historical data from prior reviews, matching the profile of a reviewer based on their past decisions and subject matter expertise. This setting replicates the reviews collected in real-world datasets.
- **Neural Network-Optimized Persona**: In this approach, a small neural network is trained to select reviewer personas that maximize the paper acceptance rate.

Results (Table 7) depict that *Neural Network-Optimized Persona* and *historical data initialization* produce the most consistent results, achieving a 0.74 and 0.70 on ICLR 23, respectively. This suggests that further tuning of the network could reduce over-acceptance of borderline papers, while maintaining a strong acceptance rate for high-quality submissions. On the other hand, the *random selection* method, while less accurate, performs at 0.64. For the *random selection* setting, inconsistencies arise due to misaligned reviewer traits, such as focus areas or strictness levels, which can lead to decisions that diverge significantly from those of real reviewers. These findings underscore the importance of the profile module, and its impact on achieving human-like evaluations.

A.4.4 Effect of Reviewer Expertise Level

The effect of reviewer expertise on the acceptance likelihood and feedback quality is unclear. Thus, we investigate how varying levels of simulated reviewer expertise impact the quality and consistency of reviews. We set the persona of reviewers with the following expertise levels: a *Novice Reviewer* with limited knowledge, an *Intermediate Reviewer* with general familiarity with the field, and an *Expert Reviewer* with deep expertise in the research domain. As a baseline, we also report the results of GAR that assigns the expertise level based on genuine values from OpenReview.

The results, summarized in Table 8, reveal that reviewers with higher expertise levels consistently outperformed those with lower expertise. That is, expert reviewers achieve the highest accuracy and consistency, approaching the level expected of human experts. Notably, novice and intermediate reviewers also produced reasonably accurate assessments, but their performance lagged behind that of expert-level reviewers. This suggests that emulating expertise levels in simulated reviewers improves the quality of automated reviews, supporting the use of calibrated expertise to enhance the reliability and value of LLM-based reviews.

A.4.5 Review Scores Across Different LLMs

This experiment examines the alignment between the LLM-based reviewer and human reviewers on key review criteria: *soundness*, *presentation*, and *contribution*. A subset of 1,000 papers from ICLR 2023 was selected, with both human and LLM-based reviewers providing scores for each criterion

Methods	NeurIPS		ICLR 22		ICLR 23	
	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑
Novice Reviewers	0.57±0.05	0.55±0.05	0.61±0.04	0.59±0.05	0.59±0.04	0.57±0.05
Novice Reviewers$^{>}$	0.59±0.05	0.56±0.05	0.63±0.04	0.60±0.05	0.61±0.04	0.58±0.05
Intermediate Reviewers	0.60±0.05	0.58±0.05	0.64±0.04	0.62±0.05	0.62±0.04	0.60±0.05
Intermediate Reviewers$^{>}$	0.62±0.05	0.59±0.05	0.66±0.04	0.64±0.05	0.64±0.04	0.62±0.05
Expert Reviewers	0.62±0.07	0.60±0.04	0.65±0.05	0.62±0.04	0.64±0.05	0.61±0.05
Expert Reviewers$^{>}$	0.66±0.05	0.61±0.03	0.68±0.05	0.63±0.05	0.67±0.05	0.62±0.04
GAR	0.64±0.05	0.61±0.04	0.68±0.03	0.66±0.05	0.66±0.04	0.60±0.04
GAR$^{>}$	**0.68±0.05**	**0.62±0.05**	**0.71±0.04**	**0.67±0.06**	**0.70±0.05**	**0.69±0.05**

Table 8: Performance for predicting the acceptance of manuscripts with varying levels of review expertise.

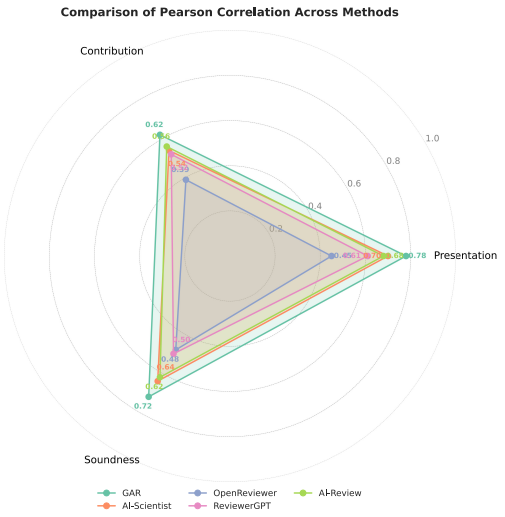

Figure 3: Alignment Between Human and LLM Reviewer Scores.

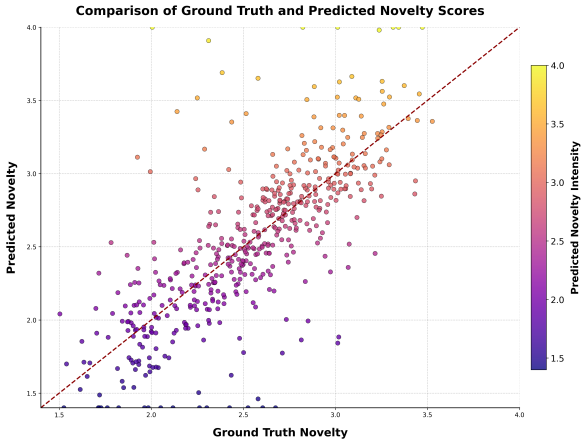

Figure 4: Comparison between predicted and ground truth novelty scores across 500 papers. Each dot represents a paper's score. The color gradient denotes the intensity of predicted novelty.

on a scale from 1 (Poor) to 4 (Excellent). Pearson correlation coefficients were calculated to measure the alignment of scores between human and LLM reviewers.

As illustrated in Figure 3, GAR demonstrates high correlation coefficients across all criteria, suggesting that the LLM-based reviewer aligns closely with human reviewers. This strong alignment, especially on aspects of soundness and presentation, highlights the LLM's ability to approximate human assessment. While the correlation for the contribution attribute is comparatively lower, our approach still surpasses prior work. This may be attributed to the inherent challenge of assessing a paper's novelty based solely on limited contextual information, whereas human reviewers benefit from extensive field-specific expertise and years of experience.

A.4.6 Novelty Score

We now assess the alignment between predicted and *ground truth* novelty scores derived from historical review data. Figure 4 presents a scatter plot comparing predictions and true novelty values. The diagonal line ($y = x$) indicates perfect alignment, while deviations highlight prediction errors. Results show a moderate correlation, with predictions closely following ground truth in the mid-range (2.5–2.9). However, deviations increase at the extremes: the model *overestimates* low-novelty papers (1.4–2.0) and *underestimates* highly novel papers (3.5–4.0), suggesting limitations in capturing novelty of papers with very low or high novelty scores.

A.4.7 Main Concerns in Reviews

To understand whether some aspects of reviews are more/less likely to be discussed by agent and human reviewers, we analyze 11 aspects of comments. Human annotation was performed a randomly sampled subset of feedback, following established research in machine learning peer review (Birhane et al., 2022; Smith et al., 2022). Figure 5 displays the relative frequency of each feedback aspect for

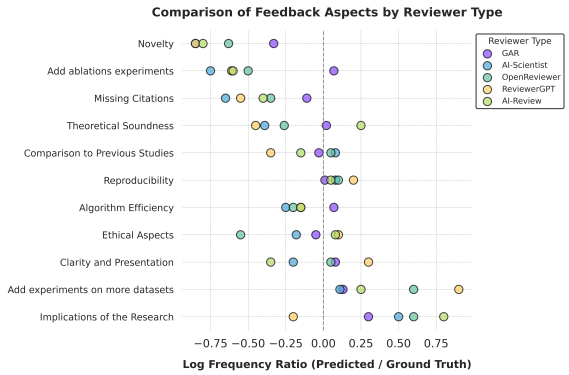

Figure 5: Relative frequency of feedback aspects by different types of reviewers.

all six reviewer types: GAR, AI-Scientist, OpenReviewer, ReviewerGPT, and AI-Review. Taking advantage of its graph-based representation, GAR aligns more closely with human reviewers than prior work, particularly in technical domains such as experiments and results. While AI-Scientist and OpenReviewer frequently highlight experimental recommendations, GAR offers a more balanced assessment.

A.4.8 Human Likeness Across Foundation Models

	NeurIPS	ICLR 22	ICLR 23
GPT-4o	**3.91 ± 0.10**	**4.08 ± 0.10**	**4.11 ± 0.10**
Mistral-7b Instruct	3.59 ± 0.08	3.67 ± 0.10	3.68 ± 0.11
Llama-3.1 (8b)	3.33 ± 0.08	3.64 ± 0.11	3.64 ± 0.12
Llama-3.1 (70b)	3.66 ± 0.09	3.63 ± 0.10	3.73 ± 0.07
GPT-4o-mini	3.89 ± 0.11	**4.02 ± 0.10**	**3.99 ± 0.09**

Table 9: Human likeness with different types of foundation LLMs.

We assess the human-likeness of reviews generated by various foundation models, including GPT-4o, GPT-4o-mini, Mistral-7b Instruct, Llama-3.1 (8b), and Llama-3.1 (70b). The results, shown in Table 9, demonstrate that GPT-4o achieves the highest overall scores across all datasets, with a score of 4.11 ± 0.10 on ICLR 2023. This highlights GPT-4o's ability to closely mimic human review styles. Overall, GPT-4o and GPT-4o-mini emerge as the most human-like in their feedback across datasets, particularly on ICLR 2023.

A.4.9 Acceptance Prediction Across Foundation Models

We now seek to evaluate the performance of our methodology using various foundation models on the acceptance prediction task. Specifically, we compare the results obtained by employing GPT-4o-mini, GPT-4o, Mistral-7b Instruct, Llama-3.1 (8b) and Llama-3.1 (70b). The results, presented in Table 10, demonstrate that the performance of GAR is generally robust across different foundation models. While GPT-4o exhibits significantly higher F1-score score (t-test $p < 0.05$), GPT-4o-mini achieves similar performance but with a lower inference time. Mistral-7b Instruct also performs reasonably well on the ICLR dataset. Among the smaller models, Mistral-7b Instruct$^{>}$ shows notable improvements, making it a competitive option for resource-constrained applications. However, Llama-3.1 models, particularly the 70b variant, demonstrate only modest gains despite their larger size, indicating diminishing returns for increased model complexity in this specific task, similarly to results obtained in Section A.4.8.

A.4.10 Impact of the Number of Reviewers on Acceptance Prediction

We investigate the impact of the number of reviewers on acceptance prediction by varying the number of available reviews from 1 to 5. To ensure consistency, ground truth labels are derived exclusively from papers that received five reviews. This setup allows us to assess how additional reviewer perspectives influence prediction performance.

Figure 6 reports the results. Increasing the number of reviewers significantly enhances performance, particularly from 1 to 3 reviewers, indicating that the early integration of multiple perspectives yields substantial improvements. However, beyond three reviewers, the performance gains diminish, suggesting that additional reviews contribute to robustness but with limited marginal benefits. This aligns with prior findings in Section 4.3, where the model's ability to integrate reviewer feedback played a key role in decision accuracy.

A.5 Review Score Alignment

To evaluate the fidelity of each reviewer model in replicating real-world review judgments, we measure the agreement between model-generated review scores and ground-truth scores obtained from geniune reviewers. We report quadratic-weighted Cohen's κ, a standard metric for ordinal score alignment, across three datasets (NeurIPS 2023, ICLR 2022, ICLR 2023). Results (Table 11) show that GAR achieves near-human alignment, outperforming prior LLM-based baselines on all datasets. This

Methods	NeurIPS		ICLR 22		ICLR 23	
	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑
GPT-4o	**0.73 ± 0.03**	**0.71 ± 0.03**	**0.75 ± 0.03**	**0.72 ± 0.03**	**0.74 ± 0.03**	**0.73 ± 0.03**
GPT-4o$^{>}$	0.70 ± 0.03	0.68 ± 0.03	0.73 ± 0.03	0.70 ± 0.03	0.72 ± 0.03	0.71 ± 0.03
Mistral-7b Instruct	0.62 ± 0.04	0.60 ± 0.04	0.64 ± 0.04	0.62 ± 0.05	0.65 ± 0.04	0.63 ± 0.05
Mistral-7b Instruct$^{>}$	0.66 ± 0.05	0.64 ± 0.04	0.68 ± 0.04	0.66 ± 0.04	0.69 ± 0.05	0.67 ± 0.05
Llama-3.1 (8b)	0.60 ± 0.06	0.58 ± 0.05	0.62 ± 0.05	0.60 ± 0.05	0.63 ± 0.04	0.61 ± 0.05
Llama-3.1 (8b)$^{>}$	0.61 ± 0.04	0.59 ± 0.05	0.63 ± 0.05	0.61 ± 0.05	0.64 ± 0.04	0.62 ± 0.05
Llama-3.1 (70b)	0.63 ± 0.03	0.61 ± 0.04	0.65 ± 0.07	0.63 ± 0.04	0.66 ± 0.03	0.64 ± 0.07
Llama-3.1 (70b)$^{>}$	0.67 ± 0.04	0.65 ± 0.04	0.69 ± 0.04	0.67 ± 0.03	0.70 ± 0.05	0.68 ± 0.06
GPT-4o-mini	0.64±0.05	0.61±0.04	0.68±0.03	0.66±0.05	0.66±0.04	0.60±0.04
GPT-4o-mini$^{>}$	0.68±0.05	0.62±0.05	0.71±0.04	0.67±0.06	0.70±0.05	0.69±0.05

Table 10: Performance comparison of GAR and baselines on three datasets, each consisting of 1,000 papers. Results are presented for different foundation models.

	NeurIPS	ICLR 22	ICLR 23
	κ	κ	κ
AI-Scientist	0.50 ± 0.04	0.56 ± 0.04	0.52 ± 0.05
OpenReviewer	0.35 ± 0.04	0.37 ± 0.04	0.35 ± 0.04
ReviewerGPT	0.47 ± 0.05	0.51 ± 0.05	0.48 ± 0.05
AI-Review	0.48 ± 0.05	0.53 ± 0.05	0.54 ± 0.05
GAR	0.61 ± 0.04	0.63 ± 0.03	0.62 ± 0.04

Table 11: Alignment between synthetic reviewer scores and OpenReview ground-truth ratings, measured with quadratic-weighted Cohen's κ (higher = better). Bars show relative performance; medals mark the top-3 per dataset. Mean ± std over three seeds.

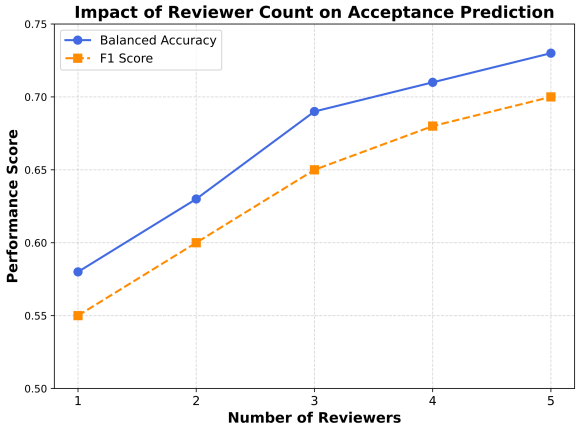

Figure 6: Effect of the number of reviewers on acceptance prediction.

confirms GAR's effectiveness in producing reviews that closely match actual reviewer scoring patterns.

A.6 Accuracy of Profile Attributes

Among the profile attributes, we focus on **expertise level**, as a reliable ground-truth can be retrieved from OpenReview. Both NeurIPS and ICLR require reviewers to self-report their confidence or expertise level for each review. We evaluate the accuracy of GAR's inferred expertise by comparing predicted reviewer confidence to the actual self-reported values, using quadratic-weighted Cohen's κ as the primary metric. On NeurIPS 2023, GAR attains a κ of 0.65 ± 0.05; for ICLR 2022, 0.68 ± 0.08; and for ICLR 2023, 0.65 ± 0.06. Macro-F1 scores range from 0.51 to 0.55. These results indicate that GAR can reliably recover reviewer expertise from review text and interaction history, achieving substantial agreement with real scores. While we expect similar trends for other profile attributes, their systematic evaluation is left to future work.

Phase	Tokens (Input/Output)	Estimated Cost (USD)
Graph Construction	2,500 / 400	0.00037
Novelty Assessment	1,000 / 300	0.00023
Review Generation	6,000–9,000 / 2,000–3,000	0.00060–0.00390
Meta-Review	1,500 / 400	0.00047
Total	11,000–14,000 / 3,100–4,100	**0.00094–0.00572**

Table 12: Estimated token consumption and cost per paper.

A.7 Running Time Analysis

We evaluate the computational efficiency of GAR by measuring the time required to process 1,000 papers under sequential and parallelized execution. When running sequentially, GAR requires 10.8 hours to process 1,000 papers due to additional LLM calls for multi-round refinement and community-based retrieval. Parallel execution significantly reduces runtime to 0.62 hours, demonstrating the scalability of the approach.

A.8 Cost Analysis

Table 12 summarizes the estimated token usage per paper. The cost scales linearly with the number of reviewers and refinement rounds. For large-scale evaluations, processing 1,000 papers incurs an estimated cost of $0.94–$5.72, assuming the same number of reviewers and refinement rounds. For 10,000 papers, the estimated cost similarly scales to $9.38–$58.11, and with parallel execution, GAR can complete the evaluation in under 9 hours, well within typical conference deadlines. Optimizing inference through batch processing and caching can further reduce costs. Compared to human peer review, which is time-intensive and costly, GAR provides an efficient and scalable alternative while maintaining review depth and consistency.

A.9 Bias Amplification and Mitigation

Institutional bias is a critical concern in peer review. To investigate this, we measured acceptance rates (ICLR 2023) conditioned on whether at least one author is affiliated with a QS-2025 top-50 university. Table 13 compares the distribution of submissions, real-world accepted papers, and GAR reviewer-agent decisions. GAR's acceptance share for top-50 institutions (69.2%) is substantially closer to the submission pool baseline (66.1%) than the real-world outcome (72.2%), while maintaining similar overall acceptance. This suggests that GAR may reduce, rather than amplify, institutional bias.

Split	≥1 Top-50	Non-Top-50	Total
Overall submission pool	66.1%	33.9%	3,793
Real-world accepted	72.2%	27.8%	1,572
GAR reviewer-agent (accepted)	69.2%	30.8%	—

Table 13: Acceptance rates conditioned on institutional affiliation.

A.10 Impact of Reviewer Personas

To directly assess the contribution of reviewer personas, we introduce a no-persona baseline, where agents perform reviews without any persona initialization. Table 14 reports results across NeurIPS, ICLR'22, and ICLR'23. Relative to the no-persona baseline, historical personas consistently improve balanced accuracy and F1 scores, demonstrating that persona modeling contributes measurable gains in predictive performance. Historical personas consistently yield improvements in balanced accuracy and F1 over no-persona variants, confirming that persona modeling contributes to GAR's performance gains.

A.11 Discussion

In this work, we present GAR, one of the first framework for simulating the peer review process through the use of LLM-empowered agents. GAR agents autonomously analyze the manuscripts, evaluate their content, provide feedback, and predict acceptance outcomes. This end-to-end framework integrates stages of novelty assessment, multi-round review, and meta-review, aiming to replicate genuine reviewers in an cost-efficient and scalable manner. As a demonstration, the proposed method has been applied to major machine learning conferences, showcasing its potential to provide human-like feedback and determine which papers meet the quality standards for publication.

We also acknowledge that our method has certain limitations. One remaining challenge is identifying genuinely groundbreaking or paradigm-shifting ideas. GAR presents a novelty module that leverages external knowledge to detect innovative contributions at the paper-level. However, future work should focus on equipping synthetic reviewers with the ability to recognize novelty at a more nuanced level. This may include leveraging the knowledge graph structure of manuscripts to assess paper novelty at a community level, or using citation embeddings to capture shifts in research topics and trends (Shibayama et al., 2021).

Our experiments are limited to machine learn-

Methods	NeurIPS		ICLR 22		ICLR 23	
	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑	Balanced Acc. ↑	F1 Score ↑
GAR (random persona)	0.59±0.05	0.57±0.05	0.60±0.07	0.68±0.06	0.61±0.03	0.61±0.06
GAR$^{>}$ (random persona)	0.62±0.05	0.59±0.04	0.63±0.04	0.69±0.05	0.64±0.06	0.65±0.06
GAR (historical persona)	0.65±0.05	0.60±0.04	0.65±0.06	0.64±0.04	0.69±0.06	0.66±0.05
GAR$^{>}$ (historical persona)	0.68±0.05	0.62±0.05	0.71±0.04	0.67±0.06	0.70±0.05	0.69±0.05
GAR (no persona)	0.56±0.05	0.55±0.05	0.60±0.05	0.66±0.05	0.59±0.04	0.58±0.05
GAR$^{>}$ (no persona)	0.60±0.05	0.57±0.05	0.63±0.04	0.67±0.05	0.62±0.05	0.63±0.05
GAR	0.64±0.05	0.61±0.04	0.68±0.03	0.66±0.05	0.66±0.04	0.60±0.04
GAR$^{>}$	**0.68±0.05**	**0.62±0.05**	**0.71±0.04**	**0.67±0.06**	**0.70±0.05**	**0.69±0.05**

Table 14: Ablation on reviewer personas across NeurIPS, ICLR'22, and ICLR'23. **Bold**: best results; underline: second-best. Results are averaged across 3 seeds.

ing conferences (ICLR, NeurIPS), and we recognize that peer review practices vary widely across disciplines. As such, GAR's generalizability beyond these domains remains to be established. Future work will involve adapting and validating our framework in fields with different review norms, such as mathematics or experimental sciences, which present unique challenges, including the evaluation of mathematical proofs or detailed experimental protocols.

Practical deployment of GAR in real-world peer review systems presents several implementation challenges. Ensuring effective oversight by editors or chairs is essential for maintaining review quality and accountability, necessitating transparent controls and intervention mechanisms. Additionally, handling sensitive information requires confidentiality and intellectual property safeguards. Addressing these challenges is critical for practical adoption and widespread impact.

Despite efforts to reduce bias, AI models like GAR are not immune to inherent biases present in training data, which can impact the evaluation process and potentially disadvantage certain research fields or authors (Gallegos et al., 2024). Establishing clear guidelines for ethical AI evaluation, incorporating fairness checks, regular audits, and opportunities for human oversight will be critical for maintaining trust in AI-driven review processes (Haffar et al., 2019). Furthermore, we must question: *Are we certain that these papers are not already part of the LLMs' training corpus?* If such overlap exists, it could inadvertently introduce bias, as the model may demonstrate familiarity with the content, concepts, or style of certain papers, providing an unfair advantage or skewing evaluations. This issue is particularly pronounced for widely circulated preprints or seminal works that are likely to have influenced the training datasets of LLMs. Addressing this challenge presents a promising direction for future research.