DREAM: Domain-Retrieved Evidence and Multi-Agent Reasoning for Structured Paper Evaluation

Anonymous ACL submission

Abstract

With the rapid growth of global academic output, traditional peer review faces mounting challenges such as reviewer shortages and inconsistent quality. While large language models (LLMs) show promise for automation but still trail human experts in accuracy, interpretability, and bias control.

To overcome these issues, we propose a modular AI-assisted review system designed to enhance the accuracy, consistency, and professionalism of automated reviews. The system integrates three key components: (1) a **Memory Module** for reflective learning, which records past errors and improves decision stability; (2) **Retrieval-Augmented Generation (RAG)** to incorporate cross-domain factual knowledge; and (3) a **Multi-Agent System (MAS)** that delegates subtaskssuch as readability, methodology, and contributionto specialized agents, with a final agent consolidating the outputs.

Using `ECNU-SEA/SEA-E` as the base model (F1 = 0.700), adding the Memory Module boosts performance to 0.771 by significantly improving true positive (TP) rates. The MAS alone achieves 0.750, while fine-tuning each agent with QLoRA increases it to 0.786. Combining MAS and Memory further lifts performance to 0.801, demonstrating the systems synergy, scalability, and robustness.

1 Introduction

Peer review is vital to scientific publishing, ensuring research quality and credibility through expert evaluations (Johnson et al., 2018; Price and Flach, 2017). However, growing submission volume and reviewer inconsistency pose major challenges (Bornmann and Mutz, 2015; Hanson et al., 2024; Abdalla et al., 2023; Stelmakh et al., 2021), prompting interest in automated systems to improve scalability and consistency.

Despite advancements in language generation, AI systems often hallucinate facts, struggle with interpretability, and reflect training biases (Jiang et al., 2023; Liu and Shah, 2023; Jiang et al., 2024; Guo et al., 2023a). Current models remain less reliable than humans, especially on complex submissions[1]. Addressing accuracy, fairness, and transparency is crucial for future AI-assisted review systems (Kuznetsov et al., 2022).

We adopt `ECNU-SEA/SEA-E`[2] as the baseline model and enhance it with: (1) **Reflective Memory**, which reuses prior mistakes to improve future reviews; (2) **RAG**, which provides grounded, cross-domain knowledge; and (3) a modular **Multi-Agent System** (MAS), with specialized agents for Summary, Readability, Methodology, and Contribution, coordinated by a Final Judgment Agent.

Our results show that Reflective Memory alone boosts the baseline F1 score from 0.700 to 0.771 by improving true rejection rates. The MAS further improves decision quality (F1 = 0.750), and combining QLoRA with Reflective Memory achieves a peak F1 of 0.801. These findings highlight the complementary value of our proposed components in enhancing automated peer review robustness.

2 Related Work

With the rise of generative AI, **Automated Scholarly Paper Review (ASPR)** has emerged as a key field linking NLP and academic publishing. This section reviews datasets, retrieval-augmented generation (RAG), and multi-agent strategies in addressing persistent peer-review issues of transparency, fairness, and consistency.

[1] `https://statevalbase.fudan.edu.cn/info/1050/1886.htm`

[2] `https://huggingface.co/ECNU-SEA/SEA-E`

2.1 Peer Review

Peer review underpins academic publishing by evaluating **validity** and **suitability** of submissions[3]. Yet, issues like anonymity-driven bias, inconsistent standards, and review quality variability persist[4], motivating reforms in transparency and standardization. To support automation, understanding existing peer review datasets is essential. Table 1 outlines key datasets used in this study, highlighting structure, annotation quality, and generative suitability.

PeerRead (Kang et al., 2018) initiated public peer review datasets but is limited in scope. **NLPeer** (Dycke et al., 2023) unifies review formats across domains, while **SubstanReview** (Guo et al., 2023b) introduces claim-evidence annotations. **RR-MCQ** (Zhou et al., 2024) focuses on revision comprehension, and **ReviewCritique** (Du et al., 2024) compares LLM- and human-generated reviews.

Despite their value, these datasets lack domain diversity, consistent annotations, and expert feedback. To address this, our system incorporates structured tasks and prompts aligned with agent-level goals, promoting critique quality through principled dataset design.

2.2 Architectures for Automated Peer Review

SEA (Yu et al., 2024) presents a modular framework with three subsystems: **SEA-S** for standardizing inputs, **SEA-E** for generating reviews, and **SEA-A** for analyzing review quality. Its architecture guides our system in formatting, generation logic, and end-to-end modular integration.

3 Methodology

To address both efficiency and quality challenges in peer review, we propose a modular automated review system that integrates knowledge retrieval, diverse generation strategies, and standardized output formatting. The overall architecture is illustrated in Figure 1, and consists of five main modules:

1. **Input and Keyphrase Extraction:** Given a target paper (in PDF format; see Section 3.1.1), we extract the top-m keyphrases from its title and abstract using a Keyphrase Generation (KPG) model. These serve as semantic queries for subsequent retrieval.

2. **Retrieval Module:** As detailed in Section 3.3, the system performs semantic search over external knowledge sources (Section 3.1.2) to retrieve relevant literature. The retrieved content is summarized to supply auxiliary contextsuch as background, limitations, and related workfor downstream review generation.

3. **Review Strategy Module:** Based on the input paper and retrieved knowledge, the system applies one of several generation strategies (Section 3.4):

   (i) **Reflective Memory** (Section 3.4.2) integrates past model errors as reminders within prompts to reduce repeated mistakes;

   (ii) **Multi-Agent Review** (Section 3.4.1) divides the task among agents focused on Readability, Methodology, and Contribution.

4. **Standardization Module:** Generated reviews are formatted (Section 3.2) to ensure consistent structure, labeling, and fluency.

3.1 Data Preparation

We constructed two datasets (Figure 2): (1) a **Multiple Reviews Dataset** for model training/evaluation, and (2) an **Additional Knowledge Corpus** for retrieval in the RAG module. These datasets are non-overlapping and serve complementary purposes.

3.1.1 Multiple Reviews Dataset

Peer review data is collected from OpenReview, including multiple comments, meta-reviews, and decisions. The dataset spans accepted and rejected papers across ICLR and NeurIPS 2024, as detailed in Table 2, and aligns with our agent-based review dimensions.

3.1.2 Additional Knowledge

We compiled a cross-domain corpus from ACL Anthology and ScienceDirect (NLP, ML, legal tech). Documents were cleaned, normalized, and embedded for retrieval. Only papers from 2016 onward were included. Table 3 summarizes the data by source and subdomain.

[3] https://jams.pub/blog/post/why-is-peer-review-important, https://www.publisso.de/en/advice/publishing-advice-faqs/peer-review

[4] https://www.enago.tw/academy/journal-editors-deal-with-mixed-reviewer-comments/

Table 1: Representative datasets for automated review generation

Dataset	Time Span	Papers	Reviews	Avg. Reviews/Paper	Limitations
PeerRead	2007–2017	14,700	10,700	2.12	Outdated, subjective, domain-biased
NLPeer	2012–2022	5,672	11,515	1.72	Domain-limited, annotation bias
SubstanReview	2012–2022	550	550	1.00	Narrow scope, domain-limited
RR-MCQ	2022–2023	14	55	3.93	Reliability, limited scale
ReviewCritique	2020–2023	100	380	3.80	Domain knowledge gap, low diversity
NeurIPS-23	2023	3,368	15,027	4.46	Subjectivity, bias from open reviews

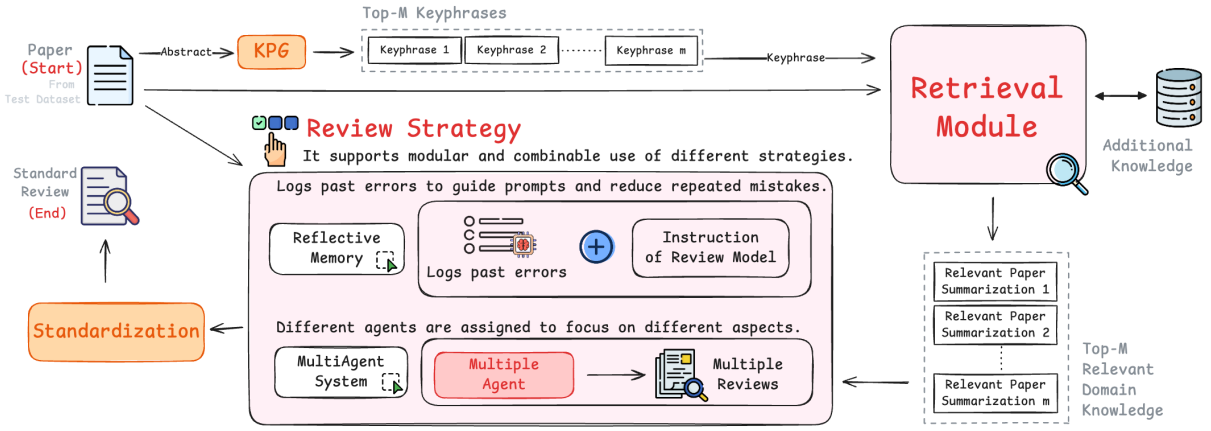

Figure 1: Overall architecture of the proposed automated review system, including retrieval, generation strategies, and standardization.

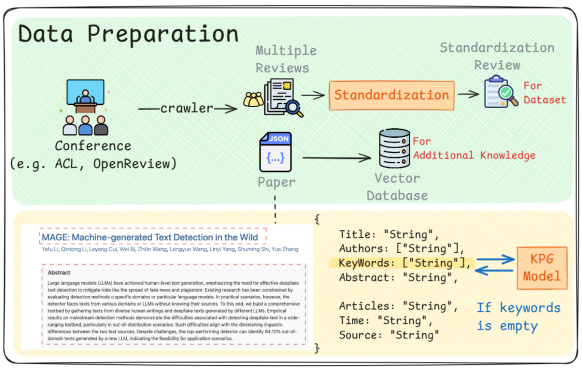

Figure 2: Data preparation pipeline. Structured review data is used for model training, while external papers are stored in a vector database for retrieval-based augmentation.

Table 2: Statistics of the collected review dataset

Conference	Year	Accepted	Rejected	Total	Acceptance Rate
ICLR	2024	2,258	3,441	5,699	39.6%
NeurIPS	2024	4,024	200	4,224	95.3%

Table 3: External knowledge dataset statistics by source and subdomain

Source	Subdomain / Type	Count
ACL	ACL 2023 Short Papers	155
ACL	ACL 2023 Long Papers	827
ACL	Findings of EMNLP 2024	969
ACL	NAACL 2024 Long Papers	470
ACL	NAACL 2024 Short Papers	72
ScienceDirect	NLP Domain	114
ScienceDirect	Legal Tech / CLSR	627
Total		**3,234**

3.1.3 Evaluation Metrics

We evaluate the quality of generated reviews using three types of metrics, each aligned with a specific review task: (1) text generation, (2) numeric score prediction, and (3) binary decision classification. Figure 3 summarizes the mapping between output components and their corresponding evaluation methods.

For textual fields like `Summary`, `Strengths`, `Weaknesses`, and `Questions`, we evaluate lexical and semantic fidelity using BLEU, ROUGE, and BERTScore.

For numerical ratings fields like `Soundness`, `Presentation`, `Contribution`, `Rating`, RMSE is computed against human annotations, while binary decisions (`Accept` vs. `Reject`) are evaluated by F1 score for classification balance.

3.2 Standardization Module

Peer reviews for a single paper often vary in tone, structure, and judgment due to differences in reviewer expertise and writing style, sometimes

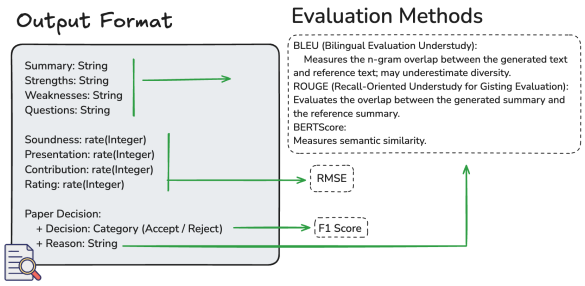

Figure 3: Evaluation framework. Text segments are assessed using BLEU, ROUGE, and BERTScore; rating predictions are evaluated via RMSE; final decisions are measured by F1 score.

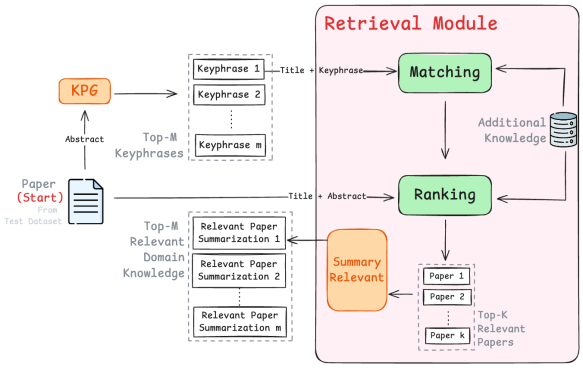

Figure 4: Architecture of the two-stage retrieval module.

even presenting conflicting opinions. These inconsistencies increase the cognitive burden on editors and may compromise fairness.

To address this, we introduce a **Standardization Module** that merges multiple reviews into a coherent, structured summary. It removes redundancy, resolves contradictions, and emphasizes key strengths and weaknesses.

This module is implemented using the SEA-S model[5], which takes all original reviews and a standardization prompt as input, producing a unified, decision-ready review. The result is a concise and balanced summary that supports fair and interpretable evaluation (Appendix A.1).

3.3 Retrieval Module

The Retrieval Module enhances review generation with domain-specific knowledge by embedding input papers as semantic vectors and retrieving top matches from a cross-domain database (NLP, ML, legal tech; see Section 3.1.2). We implement a Two-Stage Retrieval Strategy a hierarchical framework illustrated in Figure 4:

1. **Matching Stage**: Generate a vector $\mathbf{v}_{\text{title_keywords}}$ from the papers title and keywords to retrieve domain-relevant candidates.
2. **Ranking Stage**: Refine candidates using a more detailed vector $\mathbf{v}_{\text{title_abstract}}$ derived from the title and abstract.

This layered design enhances both efficiency and quality: the first stage filters out irrelevant content, while the second stage ensures high semantic alignment. The final background summary serves as a grounded, context-aware input for review generation, improving factual accuracy and cross-domain robustness.

[5] https://huggingface.co/ECNU-SEA/SEA-S

3.4 Review Strategy

The choice of review strategy critically affects the quality and traceability of automated feedback. We compare a **baseline single-model approach** with our **composite strategies**, which flexibly combine **Multi-Agent Systems (MAS)**, **Reflective Memory**, and **Paper Compression**.

3.4.1 Multi-Agent System Architecture

Based on rejection reason analysis, we design a modular Multi-Agent System (MAS) targeting clarity, novelty, and methodological rigor to improve interpretability and task specialization in review generation.

As illustrated in Figure 5, each agent evaluates a specific aspect and produces structured feedback with scores. A Final Judgment Agent integrates all outputs into a coherent review and recommendation, enhancing consistency and traceability.

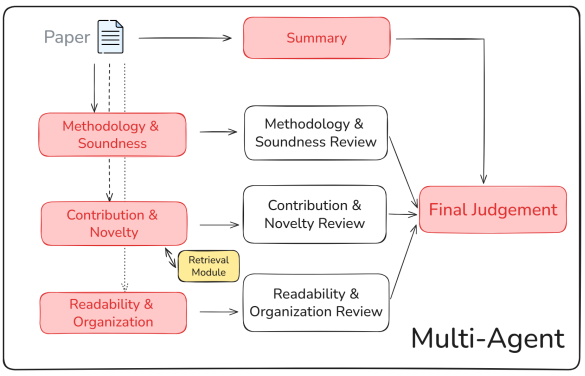

Figure 5: MAS-based review architecture. Each agent handles a dedicated review aspect; final output is composed by the Final Judgment Agent.

The system includes the following five agents:

- **Summary Agent**: Produces a neutral and concise summary of the papers motivation,

methodology, and findings to provide context for other agents (Appendix A.6).

- **Readability & Organization Agent**: Assesses language fluency, logical structure, formatting, and the clarity of visuals and tables (Appendix A.3).
- **Methodology & Soundness Agent**: Evaluates the scientific rigor and validity of the methods, including theoretical grounding, empirical design, and statistical robustness (Appendix A.4).
- **Contribution & Novelty Agent**: Uses the RAG module to compare against prior literature and assess the papers originality and substantive contributions (Appendix A.5).
- **Final Judgment Agent**: Integrates all agent outputs into a unified review and decision, ensuring coherence and completeness.

This architecture offers three key advantages:

1. improved depth and consistency across review dimensions,
2. integration of external knowledge via RAG to support factual accuracy, and
3. increased explainability and scalability through modular role-based designmaking it suitable for high-precision, structured peer review tasks.

3.4.2 Reflective Memory

Most automated review systems fail to learn from past errors, causing repeated mistakes. To address this, we propose a **Reflective Memory** module for structured self-assessment and refinement (Appendix A.7).

As shown in Figure 6, SEA-Es output is compared to human reviews. A reflective module (GPT-4o-mini) detects discrepancies and generates revision suggestions, which are distilled into reusable memory cues (e.g., Ensure control groups are included). These cues are embedded in future prompts, mimicking human learning and improving consistency and robustness over time.

4 Evaluation

This chapter evaluates the proposed modules on real-world automated review tasks. The training set includes 4,000 reviews (2,000 accepted, 2,000 rejected) for balanced classification, and another 500 reviews for testing with a 50/50 distribution.

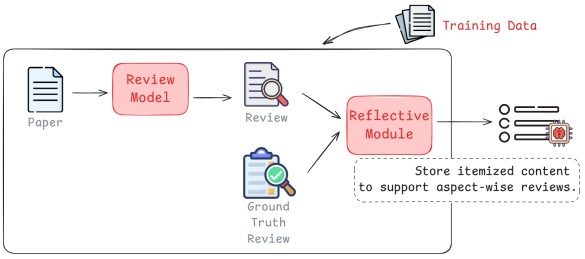

Figure 6: Reflective Memory architecture. The model compares generated and reference reviews; reflection prompts are created and saved for future reuse.

Due to space constraints, this paper focuses exclusively on the evaluation of final decision performance.

4.1 Experimental Setup

We evaluate three models**GPT-4o-mini**, **SEA-E** (baseline), and **SEA-S** (for standardization)under the following configurations:

1. **RAG-based Knowledge Augmentation**: Retrieves $k = 3$ abstracts from $m = 3$ domains (e.g., ACL, EMNLP) using extracted keywords, enriching prompts with domain-specific evidence.
2. **Multi-Agent Review Workflow**: Specialized agents assess *Methodology*, *Readability*, and *Contribution*; a final agent aggregates feedback into a unified review.
3. **Reflective Memory Mechanism**: Extracts past model errors via semantic matching and embeds correctional advice into future prompts to reduce repeated mistakes.

4.1.1 Format Validation and Output Regeneration

If the model-generated review output fails to meet the required structural formatsuch as missing sections, incorrect labels, or incomplete fieldsit is classified as a format error and excluded from immediate evaluation. To address this, we implement an automatic regeneration mechanism that re-generates the review output until it conforms to the predefined format specifications. This ensures that evaluation metrics are not distorted by occasional formatting issues, thereby maintaining fairness and consistency in the analysis.

4.2 Evaluation of Decision Accuracy

The final accept/reject decision is a critical outcome of the review process. We frame this task as a binary classification problem with two labels: `Accept` and `Reject`. The primary evaluation metric is the **Macro F1 score**, which assesses the models overall ability to identify papers that should be rejected. In addition, we analyze the proportions of True Positives (TP, correctly rejected) and True Negatives (TN, correctly accepted) to further evaluate the models classification accuracy across both classes.

4.2.1 Decision Evaluation with Reflective Memory and RAG

To evaluate the impact of auxiliary modules, we compare three review strategies using a unified output format. Table 4 reports decision performance, highlighting how different prompt contexts shape review outcomes.

- **Baseline**: Only the paper and review instruction are provided. The model relies solely on its internal knowledge, making it susceptible to hallucinations and limited coverage (Appendix A.2).
- **With RAG**: Incorporates retrieved abstracts from related papers based on the input papers topic, aiming to enrich factuality and domain-specific reasoning.
- **With Memory**: Adds a list of reflective memory entries summarizing previous review errors, providing implicit guidance to enhance reasoning and reduce bias.

Table 4: Decision classification performance across different strategies. Note: TP = True Positive (Correct Reject); TN = True Negative (Correct Accept)

Model	Strategy	Success	F1 Score	TP (%)	TN (%)
SEA	Baseline	356	0.700	51.4	90.8
SEA	With RAG	355	0.696	49.4	92.4
SEA	With Memory	**387**	**0.771**	65.5	89.2
FT-QLoRA	Baseline	360	0.711	90.0	54.2
FT-QLoRA	With RAG	**372**	**0.733**	95.2	53.8
FT-QLoRA	With Memory	348	0.684	89.6	49.8

For the SEA model, the baseline achieved high acceptance precision (TN = 90.8%) but low rejection accuracy (TP = 51.4%). FT-QLoRA reversed this, with high TP (90.0%) but low TN (54.2%), indicating over-rejection. RAG amplified these tendenciesSEA became more conservative, FT-QLoRA more aggressivesuggesting RAG may reinforce model biases. In contrast, Reflective Memory improved SEAs true positive rate by 14.1% with only a 1.6% drop in true negatives, achieving the highest F1 score (0.771). However, memory offered limited gains for FT-QLoRA due to its inherent bias. Overall, SEA and FT-QLoRA benefit from different modules: SEA aligns better with memory-based correction, while FT-QLoRA gains more from retrievalhighlighting the need to tailor auxiliary components to model behavior.

4.2.2 Decision Analysis of the Multi-Agent System

This study employs a Multi-Agent System (MAS) with four specialized agents: **Summary**, **Readability**, **Methodology**, and **Contribution**, each handling a distinct review dimension. A **Final Judgment Agent** (*SEA-S*) aggregates their feedback into an overall decision.

We compare four strategies: *SEA-E baseline*, *SEA-E + Memory*, *FT-QLoRA*, and *FT-QLoRA + Memory*. All MAS setups outperform the baseline. The best result**FT-QLoRA + Memory**achieves TP = 92.4%, TN = 68.5%, and Macro F1 = **0.801**, demonstrating the effectiveness of combining fine-tuning and memory. Table 5 details metrics for all strategies and agents.

For both *SEA-E* and *FT-QLoRA*, adding **Reflective Memory** boosts TP with minimal TN loss. The best results come from combining **fine-tuning and memory**, confirming their stabilizing synergy. The **Final Judgment Agent** achieves the highest F1 (SEA-E + Memory: 0.790; FT-QLoRA + Memory: 0.801), validating multi-agent aggregation. Sub-agents play distinct roles: **Readability** is conservative, **Methodology** is balanced, and **Contribution** (with RAG) excels at TP but over-rejects. Overall, **FT-QLoRA + Memory** offers the most reliable MAS configuration.

4.2.3 Multi-Agent Ablation Study

We assess each agent**Readability (R)**, **Methodology (M)**, and **Contribution (C)**by comparing the full RMC setup with variants excluding one (**w/o C**, **w/o M**, **w/o R**). As shown in Table 6, RMC achieves the highest F1 across models, confirming agent synergy.

Removing **C** greatly lowers TP (e.g., SEA: 71.5% 47.0%), showing its role in detecting low-value papers. Excluding **M** reduces TN, weakening acceptance precision. **R** plays a minor yet supportive role in filtering poor writing.

Table 5: Performance of MAS strategies on the Final Decision Task (Accept/Reject). Note: TP = True Positive (Correct Rejection); TN = True Negative (Correct Acceptance)

Model	Method	Strategy	Success	F1 Score	TP (%)	TN (%)
SEA-E	Baseline	-	356	0.700	51.4	90.8
FT-QLoRA	Baseline	-	360	0.711	90.0	54.2
SEA-E		Readability	341	0.677	55.4	80.9
		Methodology	351	0.698	58.6	81.7
		Contribution	377	0.754	77.1	73.7
		Final Judgment	375	0.750	71.5	78.5
FT-QLoRA		Readability	374	0.748	79.5	70.1
		Methodology	388	0.776	75.9	79.3
		Contribution	327	0.622	94.8	36.3
		Final Judgment	394	**0.786**	88.4	69.3
SEA-E	with Memory	Readability	366	0.730	65.1	81.3
		Methodology	383	0.766	75.9	77.3
		Contribution	349	0.694	81.1	58.6
		Final Judgment	395	**0.790**	75.5	82.5
FT-QLoRA	with Memory	Readability	384	0.767	84.3	69.3
		Methodology	392	0.784	79.5	77.3
		Contribution	328	0.623	95.6	35.9
		Final Judgment	**402**	**0.801**	92.4	68.5

The best-performing setup**FT-QLoRA + Memory + RMC**achieves F1 = 0.801, with TP = 92.4% and TN = 68.5%, validating the benefit of all three agents for reliable review.

Table 6: Impact of different agent combinations on final review decisions. R: Readability, M: Methodology, C: Contribution.

Model		**Success**	**F1 Score**	**TP (%)**	**TN (%)**
SEA	RMC	375	**0.750**	71.5	78.5
	w/o C	338	0.661	47.0	88.1
	w/o M	357	0.711	61.0	80.9
	w/o R	367	0.732	64.7	82.1
FT-qlora	RMC	394	**0.786**	88.4	69.3
	w/o C	383	0.764	68.3	84.9
	w/o M	374	0.746	83.1	66.5
	w/o R	382	0.764	80.3	72.5
SEA-Memory	RMC	393	**0.790**	75.5	82.5
	w/o C	381	0.762	67.5	85.3
	w/o M	365	0.728	69.1	76.5
	w/o R	365	0.732	72.7	73.7
FT-qlora with Memory	RMC	402	**0.801**	92.4	68.5
	w/o C	400	0.799	75.1	84.9
	w/o M	363	0.722	85.1	60.2
	w/o R	386	0.770	85.9	68.5

5 Conclusion

We propose a modular framework for automated peer review, combining Retrieval-Augmented Generation (RAG) with a Multi-Agent System (MAS). Specialized agents assess **Readability**, **Methodology**, and **Contribution**, with a **Final Judgment Agent** aggregating their feedback. This modular design enhances interpretability and facilitates future extensibility.

Experiments confirm each modules effectiveness. The RMC agent trio achieves the best balance of precision and recall, with the **Contribution Agent** leveraging RAG to improve rejection accuracy. **Reflective Memory** boosts model stability by embedding prior errors into prompts, lowering RMSE and enhancing decision quality across configurations.

We also find that academic review requires task-specific tuning. SEA-E outperforms GPT-4o-mini in classification and long-context stability. Compression benefits GPT models but can harm SEA by discarding critical context.

The released crawling pipelinecovering ACL Anthology, ScienceDirect, and OpenReviewenables scalable dataset construction for future research. In sum, this study introduces a principled, interpretable framework for AI-assisted peer review, with future directions including ethical dimension integration, longer-context support, and human-AI collaboration.

6 Limitations and Future Work

This study improves automated academic review through RAG, Reflective Memory, and a Multi-Agent System, though several limitations remain.

Current agents cover only readability, methodology, and contribution. Future work should address ethics, reproducibility, and human-AI collaboration for more responsible evaluations. Compression improves TP but may induce rejection due to information loss. Quantitative gains alone are insufficient; future studies should incorporate qualitative analysis and attention visualization.

The Contribution Agent benefits from RAG but suffers TN drops, possibly due to misleading references. Improving retrieval relevance, assessing source credibility, and expanding corpora will enhance reliability. Lastly, general-purpose models like GPT-4o-mini lag behind task-specific SEA-E in long-context reasoning, reinforcing the need for fine-tuning.

Ethical Considerations

This work does not involve human subjects, personal data, or sensitive information. All datasets used in this study- OpenReview, ACL Anthology, and ScienceDirect- are publicly available and contain no personally identifiable information. Potential biases from pre-trained models (e.g., SEA-E, GPT-4o-mini) are mitigated through reflective memory mechanisms and cross-domain evaluation. The proposed system aims to enhance transparency and fairness in automated peer review, and we encourage responsible human oversight in its deployment.

Responsible NLP Statement

This paper follows the ACL Responsible NLP Research Checklist. Large language models (LLMs) such as GPT-4o-mini and SEA-E were used in the research pipeline for review generation and reflection, but not for manuscript writing. All model outputs were verified manually, and all datasets used are publicly available for replication and further study.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

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

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

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

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

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823843.

Tingting Jiang, Zhumo Sun, Shiting Fu, and Yan Lv. 2024. Human-ai interaction research agenda: A user-centered perspective. *Data and Information Management*, page 100078.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. *The STM Report: An Overview of Scientific and Scholarly Publishing*. International Association of Scientific, Technical and Medical Publishers, The Hague, Netherlands.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

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

Simon Price and Peter A Flach. 2017. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79.

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*, 5:1–17.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *Preprint*, arXiv:2407.12857.

Ruiyang Zhou, Lu Chen, and Kai Yu. 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)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

A Instruction

A.1 Prompt of Standardization

INSTRUCTION:
As an experienced academic paper reviewer, you are presented with different review contents for the same paper. Please analyze these contents carefully and consolidate them into a single review. The review should be organized into nine sections: Summary, Strengths, Weaknesses, Questions, Soundness, Presentation, Contribution, Rating and Paper Decision. Below is a description of each section:

Rule:
1. Summary: A summary of the paper in 100-150 words
2. Strengths/Weaknesses/Questions: The Strengths/Weaknesses/Questions of paper, which should be listed in bullet points, with each point supported by specific examples from the article where possible.
3. Soundness/Presentation/Contribution: Rate the papers Soundness/Presentation/Contribution, and match this score to the corresponding criterion from the list below and provide the result.The possible scores and their criteria are:
    1 poor
    2 fair
    3 good
    4 excellent
4. Rating: Give this paper an appropriate Rating, match this Rating to the corresponding criterion from the list below and provide the result.The Rating must be an integer, then, match this integer Rating to the corresponding criterion.The possible Ratings and their criteria are:
    1 strong reject
    2 reject, significant issues present
    3 reject, not good enough
    4 possibly reject, but has redeeming facets
    5 marginally below the acceptance threshold
    6 marginally above the acceptance threshold
    7 accept, but needs minor improvements
    8 accept, good paper
    9 strong accept, excellent work
    10 strong accept, should be highlighted at the conference
5. Paper Decision:It must include the Decision itself (Accept or Reject) and the reasons for this decision which is based on the criteria of originality, methodological soundness, significance of results, and clarity and logic of presentation.

REVIEWS:
<Reviews>

Output Format:
Summary: <Summary content>
Strengths: <Strengths result>
Weaknesses: <Weaknesses result>
Questions: <Questions result>
Soundness: <Soundness Score>
Presentation: <Presentation Score>
Contribution: <Contribution Score>
Rating: <Rating Score>
Paper Decision:
    - Decision: Accept/Reject
    - Reasons: reasons content
Ensure your feedback is objective and constructive.

A.2 Prompt of Review

INSTRUCTION:
You are an experienced academic reviewer. Evaluate the paper based on the following criteria. Structure the review into nine sections.

Paper:
<Paper Content>

Rule:
1. Summary: A summary of the paper in 100-150 words
2. Strengths/Weaknesses/Questions: The Strengths/Weaknesses/Questions of paper, which should be listed in bullet points, with each point supported by specific examples from the article where possible.
3. Soundness/Presentation/Contribution: Rate the papers Soundness/Presentation/Contribution, and match this score to the corresponding criterion from the list below and provide the result.The possible scores and their criteria are:
    1 poor
    2 fair
    3 good
    4 excellent
4. Rating: Give this paper an appropriate Rating, match this Rating to the corresponding criterion from the list below and provide the result.The Rating must be an integer, then, match this integer Rating to the corresponding criterion.The possible Ratings and their criteria are:
    1 strong reject
    2 reject, significant issues present
    3 reject, not good enough
    4 possibly reject, but has redeeming facets
    5 marginally below the acceptance threshold
    6 marginally above the acceptance threshold
    7 accept, but needs minor improvements
    8 accept, good paper
    9 strong accept, excellent work
    10 strong accept, should be highlighted at the conference
5. Paper Decision:It must include the Decision itself (Accept or Reject) and the reasons for this decision which is based on the criteria of originality, methodological soundness, significance of results, and clarity and logic of presentation.

Output Format:
Summary: <Summary content>
Strengths: <Strengths result>
Weaknesses: <Weaknesses result>
Questions: <Questions result>
Soundness: <Soundness Score>
Presentation: <Presentation Score>
Contribution: <Contribution Score>
Rating: <Rating Score>
Paper Decision:
    - Decision: Accept/Reject
    - Reasons: reasons content
Ensure your feedback is objective and constructive.

A.3 Prompt of Readability

Role
You are a skeptical reviewer focusing on readability & organization.
Instruction
Review the papers readability and structure. The default stance is **Reject**, upgrade to *Accept* only if every checkpoint shows **no major flaws**.
Scoring Procedure > Gate-by-Gate (failure in any gate Reject)
Gate-1 Structural Fatal Flaws
- Missing or out-of-order sections (e.g., no Method / Result / Conclusion).
- Inconsistent numbering or broken references for figures equations / appendices.
- **Terminology inconsistency**: key terms undefined or used inconsistently.
Gate-2 Readability Fatal Flaws
- Pervasive grammatical errors or machine-translation artifacts (3 per 1000 words).
- Figures or appendices without titles / captions, making them unintelligible on their own.
- Critical statements unclear and impacting the main results.
...
Paper Content
<Paper Content>

Rule
<Rule Content>

Output Format **Strengths:** <Strengths result>
Weaknesses: <Weaknesses result>
Questions: <Questions result>
Presentation: <Presentation Score>
Rating: <Rating Score>
Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content
Ensure your feedback is objective and constructive.

Evaluation Principles
- Gate-by-gate review; any Gate-1 or Gate-2 failure Reject.
- Each flaw must cite section or line numbers in the format: ‘(sec 2.3, line 145)‘.
- For clarity issues, include:
  1. Excerpt of the unclear sentence (if short).
  2. Description of the issue (e.g., vague term, unclear structure).
  3. Impact on understanding (e.g., claim ambiguity, result interpretation).
- Apply strict standards.
- Give credit only when fully justified.
- Only follow the output format.
- Challenge any unsupported, vague, or exaggerated claims.
- Begin with a skeptical stance: assume rejection by default, and only grant acceptance when the paper earns it through clearly supported, specific, and well-justified claims.
- Author responses are welcome; however, if the response fails to address key concerns or provide sufficient justification, the level of scrutiny should increase accordingly.

A.4 Prompt of Methodology

Role
You are a reviewer analyzing the methodology and soundness of the paper.

Instruction
Review the papers methodology and results. The default stance is **Reject**, upgrade to *Accept* only if every checkpoint shows **no major flaws**.
Scoring Procedure > Gate-by-Gate (failure in any gate Reject)
Methodology & Soundness
- Assumption Clarity: Are assumptions clearly stated and justified?
- Theoretical Soundness: Are models logically consistent and supported by formal analysis?
- Experimental Design: Is the setup appropriate, controlled, and reproducible?
- Ablation & Robustness: Are component contributions verified and sensitivity analyses conducted?
- Limitations: Are methodological risks and constraints discussed transparently?
Results & Evidence Strength
- Sufficiency of Evidence: Do the results comprehensively align with claims?
- Statistical Significance: Are statistical measures reported and interpreted correctly?
...

Paper Content:
<Paper Content>

Rule
<Rule Content>

Output Format: **Strengths:** <Strengths result>
Weaknesses: <Weaknesses result>
Questions: <Questions result>
Soundness: <Soundness Score>
Rating: <Rating Score>
Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content
Ensure your feedback is objective and constructive.

Evaluation Principles
- Apply strict standards.
- Give credit only when fully justified.
- Only follow the output format.
- Challenge any unsupported, vague, or exaggerated claims.
- Begin with a skeptical stance: assume rejection by default, and only grant acceptance when the paper earns it through clearly supported, specific, and well-justified claims.
- Author responses are welcome; however, if the response fails to address key concerns or provide sufficient justification, the level of scrutiny should increase accordingly.

A.5 Prompt of Contribution

Role
You are a reviewer analyzing the contribution and novelty of the paper.
Instruction
Review the papers originality and contributions. The default stance is **Reject**, upgrade to *Accept* only if every checkpoint shows **no major flaws**.
Scoring Procedure > Gate-by-Gate (failure in any gate Reject)
Novelty
- Does the paper introduce a new idea, method, or application?
- How does it differ meaningfully from prior work?
- Are novelty claims clearly stated and supported?
Substantive Contribution
- Does the paper provide a significant theoretical, methodological, or practical advancement?
- Are the contributions generalizable or applicable beyond narrow cases?
Positioning in Literature
- Is the contribution well-positioned in the context of related work?
- Are comparisons drawn to clearly highlight what is new or improved?
Clarity of Claims
- Are the key contributions explicitly listed or clearly emphasized?
- Is it easy for readers to understand what the paper is trying to contribute?
- Is the writing style conducive to understanding the novelty?
...

Paper Content:
<Paper Content>

Rule
<Rule Content>

Output Format:
Strengths: <Strengths result>
Weaknesses: <Weaknesses result>
Questions: <Questions result>
Contribution: <Contribution Score>
Rating: <Rating Score>
Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content
Ensure your feedback is objective and constructive.

Evaluation Principles
- Apply strict standards.
- Give credit only when fully justified.
- Only follow the output format.
- Challenge any unsupported, vague, or exaggerated claims.
- Begin with a skeptical stance: assume rejection by default, and only grant acceptance when the paper earns it through clearly supported, specific, and well-justified claims.
- Author responses are welcome; however, if the response fails to address key concerns or provide sufficient justification, the level of scrutiny should increase accordingly.

A.6 Prompt of Summary

Role
You are a reviewer summarizing the paper.

Instruction
Write a clear, concise, and accurate summary of the paper based on the full content. The summary should:
1. **Content Coverage**
   - Capture the core objective of the paper
   - Briefly describe the methodology
   - Highlight key findings or contributions
2. **Clarity & Brevity**
   - Be written in your own words (no copy-paste from abstract)
   - Stay within 100150 words
   - Use precise and professional academic language
3. **Neutral Tone**
   - Avoid subjective judgments or evaluation
   - Do not include opinions, strengths, or weaknesses

Paper Content:
<Paper Content>

Rule:
1. Summary: A summary of the paper in 100-150 words

Output Format:
Summary: <Summary content>
Ensure your feedback is objective and constructive.

A.7 Prompt of Reflective Memory

Instruction You're a professional academic reviewer. You now need to reflect on the difference between your previous review and a real review.

You will receive:
1. The original paper: [Paper]
2. Your review: [Review]
3. The real review: [Real Review]
4. Your current review memory or heuristics: [Memory]

Goal Identify key gaps or mistakes in your review compared to the real review. Focus on what **important aspects you missed, underestimated, or misjudged**, and **what should be added or updated in your memory** to improve future reviews.
Translate the reminders into more general warnings.

Output Format Respond with a concise LIST only. Each item should:
- Identify the missed/underemphasized aspect
- (Optional) Briefly suggest how to update memory

Paper <Paper Content>
Review <Review Content>
Memory <Memory Content>
Real Review <Real Review Content>
Output: use - (LIST ONLY)