MAMORX: Multi-agent Multi-Modal Scientific Review Generation with External Knowledge

Pawin Taechoyotin* **Guanchao Wang*** **Tong Zeng** **Bradley Sides**
Daniel E. Acuna+

University of Colorado Boulder

*Equal contribution.
+Corresponding author: daniel.acuna@colorado.edu

Abstract

The deluge of scientific papers has made it challenging for researchers to throughly evaluate their own and others' ideas with regards to novelty and improvements. We propose MAMORX, an automated scientific review generation system that relies on multi-modal foundation models to address this challenge. MAMORX replicates key aspects of human review by integrating attention to text, figures, and citations, along with access to external knowledge sources. Compared to previous work, it takes advantage of large context windows to significantly reduce the number of agents and the processing time needed. The system relies on structured outputs and function calling to handle figures, evaluate novelty, and build general and domain-specific knowledge bases from external scholarly search systems. To test our method, we conducted an arena-style competition between several baselines and human reviews on diverse papers from general machine learning and NLP fields, calculating Elo ratings based on human preferences. MAMORX has a high win rate against human reviews and outperforms the next-best model, a multi-agent system. We share our system[1], and discuss further applications of foundation models for scientific evaluation.

1 Introduction

AI holds a great deal of potential for accelerating knowledge production. It can help during the exploratory phases of research, such as idea generation and refinement [14], or the entire process [38]. A crucial step in producing new knowledge is reviewing whether an unpublished manuscript is worthy of publication. Previous work that has developed automated review systems shows that they are feasible [3, 43, 31] although need supervision (quality, factuality, and possible biases [43, 6, 25]). New developments in LLMs open the door to evaluate improvements in review generation. Some of these advancements include larger context windows, modalities beyond text such as images and tables, and tooling that allows them to act as agents and access information outside of their training data. In this article, we propose MAMORX, a multi-agent, multi-modal scientific review generation framework with external knowledge that takes advantage of these advances.

Recent studies have explored the potential and limitations of AI in scientific review generation. While AI can enhance efficiency and comprehensiveness in writing reviews [16, 19], significant challenges remain. These include factual inaccuracies, outdated information, and reference fabrication [4, 35].

[1]The code for our system can be found at https://github.com/sciosci/mamorx-review-system and an example implementation is running at https://rev0.ai

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

AI-generated reviews may also have higher similarity indices, increasing plagiarism risks [19, 1]. Furthermore, current AI models struggle with context understanding and generating constructive feedback [43]. In spite of these challenges, experts emphasize that AI could complement human expertise in scientific review processes [34, 19]. Ongoing research is needed to address ethical concerns, biases, and technological limitations in order to responsibly capture AI's full potential in scientific communication.

The closest work in the review generation literature tends to use expertly crafted prompts or complex agent communications. In a detailed study of the feasibility of GPT-4 reviews, Liang et al. [24] found that the overlap between GPT-4 and human feedback is comparable to the overlap between two human reviewers. However, GPT-4 often struggles with providing in-depth critique of method design and tends to focus on similar feedback, such as suggesting more datasets for experiments, making its results sometimes feel generic. To solve some of these issues, D'Arcy et al. [10] presented MARG, a multi-agent review generation system, which significantly improves the specificity and helpfulness of feedback by distributing chunks of the paper's text across specialized agents. To achieve this goal, MARG system uses a large number of agents that need heavy communication. The system is limited, however, in that it only uses the built-in knowledge of GPT-4. Recently, Lu et al. [29] introduced the "AI Scientist", a framework that automates several pieces of the scientific process, including generating research ideas, writing code, executing experiments, and producing full scientific papers. While more comprehensive, the "AI Scientist" tends to produce significantly less deep analyses, results, and, consequently, reviews.

In this article, we introduce MAMORX, a multi-agent, multi-modal scientific review generation systems with external knowledge. This system addresses several of the limitations that other systems face. Instead of using groups of agents to analyze one aspect of the paper (e.g., chunking the experiments section like in [10]), MAMORX uses a single agent for each aspect and takes advantage of the large context window of recent models. We also introduce an external tool that creates a mini-review of the foundational information for the paper being reviewed. A novelty evaluation component simulates a graph analysis, comparing the similarity of the paper being reviewed to an external scholarly search system. Finally, we introduce a multi-modal approach that is able to criticize the figures based on the context of the paper in review. In our evaluations, we find that MAMORX is able to generate reviews that are favored by humans, outperforming other baselines and human reviewers. In sum, the contributions of this paper are:

- A multi-agent multi-modal scientific review generation framework with external knowledge (MAMORX).
- An arena style system to evaluate the quality of multiple review generation systems
- A multi-modal approach capable of analyzing and critiquing figures in scientific papers.
- A novelty evaluation system using graph analysis and similarity metrics to assess the originality of the reviewed paper.
- Evaluation of MAMORX against baselines and human-generated reviews.

2 Related work

Human reviewers integrate multiple elements in their evaluation process: textual analysis, visual interpretation, citation assessment, and external knowledge application [21]. They critically analyze text to identify key points, evaluate visual data to support or challenge claims, use citations to ground arguments in existing research, and incorporate external expertise to assess novelty and relevance. This comprehensive approach should produce informative, high-quality reviews. However, challenges remain in maintaining up-to-date knowledge, especially in rapidly evolving fields. The exponential growth of science has become increasingly challenging for peer review to work effectively. The annual growth rate is approximately 8% [23, 22, 33], with some fields expanding even faster [32]. Therefore, new complementary approaches should be explored.

Automated scientific review generation has seen significant advancements in recent years. Early research explored the feasibility of automating scientific reviewing but had significant limitations such as isolation of comments and relevance [44, 3]. Content models for survey generation were introduced, paving the way for more sophisticated approaches [17]. The advent of large language

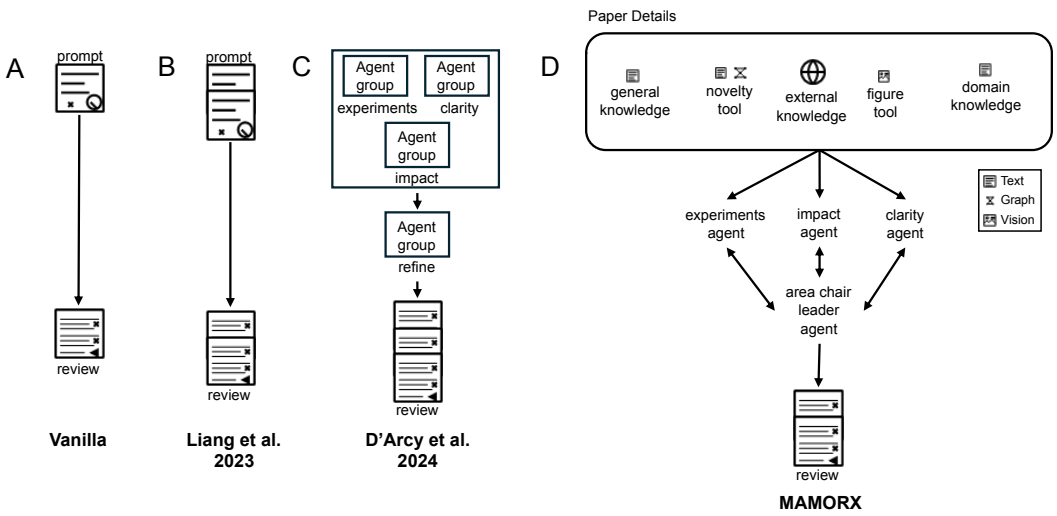

Figure 1: **Comparison of scientific review generation models:** (A) Vanilla model with a simple prompt, (B) Liang et al.'s model with a more sophisticated prompt [24], (C) D'Arcy et al.'s MARG multi-agent system [10], and (D) our proposed MAMORX (Multi-agent Multi-Modal Review Generation with External Knowledge) system. MAMORX integrates specialized agents for novelty assessment, figure criticism, and domain knowledge incorporation. It leverages multi-modal inputs (text, figures, and citation graphs) and accesses external knowledge sources. Icons denote modality types: document for text, network for citation graph, and image for figures. The system uses large context windows to reduce the number of agents and processing time compared to previous work.

models has revolutionized the field, with studies showing the power of GPT-4 in generating specific and helpful feedback compared to earlier work [10, 24]. Recent applications in academic writing have shown that while AI-generated feedback can be comprehensive, it often lacks the personalization, context-awareness, and subtle critique of human feedback [2, 30]. The use of these models in scholarly peer review presents risks such as potential biases and lack of domain-specific expertise [15].

Recent advancements in large language models and multimodality have significantly expanded the capabilities of AI systems to process and generate content across various modalities, including text, images, video, and audio [42, 5]. Models such as GPT-4V demonstrate emergent abilities such as image-based storytelling and OCR-free math reasoning [41, 5]. Multi-modal models are not new (e.g., [40, 18, 7, 39, 37]), but the quality and performance have taken off due to massive datasets and models. Despite these advancements, challenges remain in analyzing complex research content due to the high dimensionality and heterogeneity of different modalities [12], with noise and redundancy in multi-modal data affecting model performance [28]. Also, these models might require large context windows, a limitation which is sometimes solved by using agentic systems [36]. Recent advances have further enabled these tools and agents to access external tools and structured outputs, making it easy to build multi-agent systems that interact with each other and external functionality such as external knowledge [45].

3 Method

This section introduces MAMORX (Multi-Agent Multi-Modal Scientific Review Generation with External Knowledge), our proposed system for automated scientific review generation (Fig. 1). The system integrates specialized agents, leverages multi-modal inputs, and incorporates external knowledge sources to provide reviews that more closely resemble human feedback. The prompts used by the different agents and components are presented in the appendix.

3.1 Agents

Area chair (Leader Agent) The concept of a leader agent for reviews originates from D'Arcy et al. [10]. In their system, GPT-4's context window limitations restricted each worker agent to accessing only a chunk of the paper being reviewed. The leader agent's role was to coordinate workers and expert agents, enabling the integration of fragmented information through communication. In contrast, the two multi-agent systems in our study grant agents access to all textual and graphical content from a given paper. Nonetheless, we retain the area chair to facilitate communication with expert agents. This approach allows each expert agent to concentrate on specific paper aspects, preserving details like the MARG [10] system.

Novelty assessment Before the subject paper is passed into the main agent framework, it undergoes several preprocessing steps to ensure its relevance and contribution to the field. We introduce a novelty agent designed to query the Semantic Scholar [27] database, enabling dynamic access to current research in related fields. The primary role of this agent is to determine whether the subject paper is sufficiently novel before the main agent framework assesses it. Previous approaches have not incorporated external knowledge for novelty assessment, relying solely on the general knowledge embedded within foundation models.

Our process begins with an agent analyzing the subject paper's title and abstract to produce a series of three queries for the Semantic Scholar API, which are increasingly broad in scope: the first targets closely related literature, while subsequent queries expand the scope to capture relevant research from adjacent fields that the initial, specific query may miss. We compile a database of up to 30 related papers from these queries. To prevent penalizing the subject paper from building on existing work, any papers cited in the subject paper and appearing in the database are removed. This step prevents the agent from deeming the subject paper "non-novel" due to its connection to foundational work. Next, another agent compares all recommended papers to the the subject paper and individually deems them "relevant" or "irrelevant". This step removes any unrelated papers that the Semantic Scholar query might produce if given a more ambiguous query or a very niche subject paper. Finally, with this filtered dataset, a new agent completes a pairwise novelty assessment between each relevant paper and the subject paper, specifically analyzing the degree of difference between the key ideas, methods, and findings presented by the two papers. Suppose the subject paper is considered sufficiently novel compared to every other paper in the dataset. In that case, the agent will deem it "novel" and send a summarized report to the main agent framework. If the subject paper is deemed "non-novel" compared to one or more papers in the dataset, the process terminates after the entire assessment is complete, and an explanation is given for each paper that violates the novelty requirement.

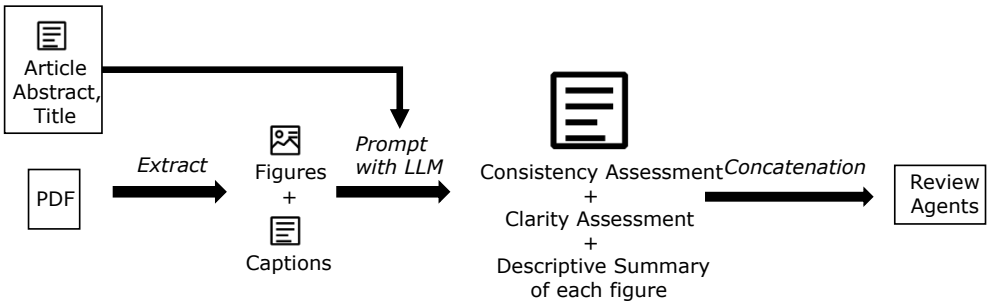

Figure 2: **Figure Critic assessment pipeline**: A pipeline depicting how the figures are extracted from the input paper before it is ready to be used by the reviewing agents. The process starts with three inputs: the PDF, title, and abstract of the paper in question. The figures and captions are then extracted from the PDF file. In the following step, it will be analyzed with the abstract and title for consistency and clarity. This step also generates a descriptive summary of each figure. Finally, the consistency and clarity assessments are concatenated with the descriptive summaries to be used for the reviewing agent crew.

Figure critic assessment Figures are a crucial part of many scientific papers. Currently, AI review systems use largely pure-text approaches, limiting the information available for review. From our initial analysis, we have found many instances where the review generated indicates a lack of explanation regarding methodologies, experiment setups, and more, even though those details are in the included figures. To address this, we propose a figure tool to address this information gap. This tool is based on the vision capabilities of large language models and aims to extract all figures within the paper under review and convey that information in text format to fit the paradigm of language models. This is done by extracting the figures and captions from the scientific paper. The extracted figure and captions are then assessed and output as a long text regarding the consistency between the title, abstract of the paper in question, and clarity assessment and description of each figure. The preprocessing of the figures is depicted in Fig. 2. We use PaperMage [26] to extract the figures and their captions from the PDF file. From our experiments, the extraction performed well with the caveat that there is a discrepancy between the number of figures and the number of captions. This also causes alignment issues where there is no simple method to associate a caption with its figure. We include all the captions when prompting the large language model to circumvent this issue. The prompting process is done independently for each figure. Apart from the figure, we include the title and abstract of the paper along with all the captions found for each prompt request. The results for each figure are then concatenated and made available to the reviewing agents.

General knowledge The notion of "general knowledge" is abstract and potentially misleading. In the MARG system, D'Arcy et al. [10] incorporated system prompts and task prompts with specific instructions for paper reviewing. These instructions target a particular field. For instance, directives like "Come up with a clear description of experiments, analyses, and ablations" and "maybe the paper includes one baseline but no others" are prevalent in machine learning (ML) but might not apply to other scientific fields. Our system adapted D'Arcy et al.[10]'s prompts, tuning them for our model's architecture. We label these specialized instruction as "general knowledge", but it's important to distinguish it from universal paper review knowledge. Instead, it refers to broad proficiency in reviewing papers within a large field — in our case, Machine Learning (ML). In contrast, the domain knowledge component (below) addresses the papers' specific sub-fields or research topics under review. For instance, while general knowledge covers broad ML concepts, domain knowledge might focus on specialized topics, such as the potential implications of AI for peer review processes.

Domain knowledge To further support the domain-specific knowledge available to the agent group, we include another preprocessing step to build a graph of cited works accessible to specific agents within the main workflow. We can access all references from the subject paper and create a graph of all content referenced by the subject paper. This is made available to agents in the workflow who require domain knowledge that pre-training did not provide. This novel inclusion enables our agents to access highly specific information, substantiating the content and depth of their reviews.

Impact agent The impact agent is an expert agent resembling the impact agent group from D'Arcy et al. [10]. In our system, a single agent was tasked to help the leader agent generate reviews concerning the impact and novelty of the paper. Our impact agent differs from the agent group in D'Arcy et al. in that it has access to the novelty tool described above.

Experiments Agent The experiments agent, similar to the experiment agent group in MARG, specializes in critiquing the paper's methodology. This expert agent focuses on aspects such as dataset quality, experimental design, and model architecture. Its role involves scrutinizing the research's technical underpinnings, ensuring a rigorous evaluation of the study's empirical base.

Clarity Agent Similar to its counterpart in MARG, the clarity agent focuses on the paper's organization, structure, and overall clarity. In addition to the LLM's inherent knowledge of grammar and writing style, this agent could assess the paper's figures and tables. This capability allows the agent to evaluate clarity from both textual and graphical perspectives, enabling it to review the paper's overall presentation comprehensively.

4 Baseline methods

In our study, the prompt-based baseline is just a review generated by prompting the full text of the paper (including sectional headings) along with a two-line task description to ask the paper to generate a review for an academic paper. To illustrate the difference between the human review and the baseline methods, we provide a stylized example in Table 1. Human reviews contain comments

from multiple categories, including general knowledge, domain knowledge, novelty, and figure criticisms. However, they tend to be more concise. Previous prompt-based methods provide more details, but they are usually not substantive and lack the depth of the human reviews [10]. Agentic systems such as MARG and ours can provide more insights by leveraging the power of multiple agents looking at different aspects of the paper. MAMORX adds other typical elements of the human review process but with deeper detail and relevance. For all our evaluations, including MAMORX, we use the Anthropic Claude 3.5 Sonnet model. MARG was originally implemented with GPT-4, but we reimplemented it with Claude 3.5 Sonnet and call it D'Arcy et al. to avoid confusion.

System	Example Review
Human Reviewer	The paper presents an interesting approach but lacks sufficient experimental validation. The proposed algorithm seems similar to Smith et al. (2022). Figure 2 is unclear and needs better labeling. The work shows some novelty but requires more comparison with recent literature. The theoretical foundation appears sound, but practical implications are not fully explored. Overall, major revisions are needed.
Prompt-based Method	The paper introduces a new method for image classification. The authors claim improved accuracy over existing methods. The methodology seems sound, but more details would be beneficial. The conclusion summarizes the main findings effectively.
Liang et al.	The proposed image classification method shows potential. The authors provide a detailed explanation of their algorithm and compare it to existing approaches. The results indicate improved performance, but more extensive experiments on diverse datasets would strengthen the claims. The paper is well-structured and written.
D'Arcy et al.	The paper presents a novel approach to image classification using a hybrid neural network architecture. The authors provide a comprehensive literature review, situating their work within the current state-of-the-art. The methodology is well-explained, with clear justifications for design choices. The experimental results are promising, showing significant improvements over baseline methods. However, the discussion of limitations could be expanded, and future work directions could be more clearly outlined. The paper's structure and writing quality are commendable, facilitating easy understanding of the complex concepts.
MAMORX (Ours)	This paper introduces a novel hybrid neural network for image classification. The authors provide a thorough literature review and clearly position their work within the field. The proposed method builds upon the work of Johnson et al. (2021) but introduces a key innovation in the attention mechanism. Analysis of the paper's references and recent publications in the field confirms the novelty of this approach. The methodology is well-explained, with clear justifications for design choices. The experimental results are promising, showing statistically significant improvements over state-of-the-art methods across multiple datasets. Figure 3 effectively illustrates the architecture of the proposed network, but the caption could be more descriptive to aid reader understanding. The discussion of limitations is comprehensive, and the proposed future work directions are both relevant and exciting. The authors' analysis of computational efficiency compared to existing methods (Table 2) is particularly insightful and adds significant value to the paper. Overall, this is a strong contribution to the field of image classification.

Table 1: Stylized examples of reviews from different systems. Color coding: blue - general knowledge comment, orange - domain knowledge comment, purple - novelty comment, red - figure criticism. Note that MAMORX incorporates external knowledge for novelty assessment and figure analysis, while other systems rely solely on their pre-trained knowledge.

Prompt-based method Our study's prompt-based baseline consists of a review generated by prompting the paper's full text, including section headings. The input is accompanied by a concise two-line task description instructing the model to produce an academic paper review. This straightforward approach serves as a comparison point for the other systems.

Advanced prompting (Liang et al., 2023) The Liang et al. [24] system, also used as a baseline in D'Arcy et al.'s study, employs a more structured approach to paper review. It consists of a detailed prompt that guides the model in producing a specific review outline. This outline includes sections on significance, novelty, and reasons for acceptance or rejection. Unlike the simple prompt-based baseline, this system provides a more comprehensive framework for evaluation. Our implementation of Liang et al. differs from the original in that we did not truncate papers exceeding the original GPT-4's token limit, potentially allowing for more complete reviews of longer papers.

D'Arcy et al, 2024 This system re-implements D'Arcy et al. [10]'s approach while utilizing the newly published version of Claude 3.5 (version = 20240620). Implemented using CrewAI, it maintains the multi-agent workflow with four key agents: leader, impact, experiments, and clarity. Like MARG, this system lacks access to graphical information, novelty assessment, domain knowledge, and external knowledge of MAMORX.

5 Evaluation

Human evaluation We collected academic papers with available peer reviews to assess our system's performance against human-generated peer reviews. Of the 30 papers in our evaluation set, 20 have corresponding human reviews. The dataset comprises ten papers from ACL 2017, sourced from the PeerRead dataset developed by Kang et al.[20] 10 papers from Advances in Neural Information Processing Systems 32 (NeurIPS 2019), which are available online. For each of these 20 papers, we

randomly selected one human review from the available set. This chosen review serves as the human benchmark for comparison in our evaluation process.

Model Evaluation with Elo Rating For our evaluation, we employed the Elo rating system, a method originally developed for calculating relative skill levels in zero-sum games [13]. This system has been adapted to evaluate language models through pairwise comparisons. For a detailed explanation of the Elo rating system and its application to language model evaluation, please see [8].

In our implementation, each reviewer (model or human) was initially assigned a rating of 1500. After each pairwise comparison, ratings were updated using the standard Elo formula:

$$R'_i = R_i + K \times (S_i - P(i \succ j)) \tag{1}$$

where $K = 32$, S_i is the actual outcome of the comparison (1 for a win, 0 for a loss), and $P(i \succ j)$ is the expected probability of reviewer i winning against reviewer j based on their current ratings.

We also computed a style-adjusted Elo rating to account for potential biases due to review formatting, using the Bradley-Terry model with covariance adjustment [11]. We use three signals of style: number of lines, number of headings, and number of lists. A preliminary analysis of these three features in isolation revealed that only the number of lines was significantly correlated with a positive vote. Still, we include all of them in the Bradley-Terry model.

Figure 3: The User Interface of the Reviewer Arena. Users can view the paper PDF and two reviews side-by-side, then evaluate them on Technical Quality, Constructiveness, Clarity, and Overall Quality

Cost analysis We did an additional analysis of input and output tokens used and the cost (at the time of publication). We ran 10 papers on MAMORX, which used an average of 1,227,850 $\pm$ 306,365 input tokens and 101,487.9 $\pm$ 20,605 output tokens. The input token's standard error of the mean is large due to the variable article sizes. At the time of writing, the Anthropic API cost was $3 per

million input tokens and $15 per million output tokens, resulting in an average of $ 5.20 ($\pm$ 0.96) per article. Finally, the timing per review is an average of 12.03 $\pm$ 0.64 minutes. A regression analysis shows that one hundred thousands tokens (input and output) add 9.7 seconds (not significant) with an intercept of 593.28 seconds.

Reviewer Arena To make it easier for users to evaluate the review models, we designed and implemented a Reviewer Arena. As shown in Figure 3, the arena interface is divided into three parts from top to bottom. The top section is an instruction panel that displays the evaluation rules to the user. The middle section is a display panel, which shows the PDF file of the paper being evaluated and the reviews written by the two Reviewers (e.g., AI models and human reviewers). The bottom section is an action panel, where users can use radio controls to compare the two reviews on four aspects: Technical Quality, Constructiveness, Clarity, and Overall Quality. It also includes buttons for navigating to the previous or next paper and submitting the user's preference. This resembles the LLM arena described in [8].

6 Results

Model	Technical Quality	Constructive-ness	Clarity	Overall Quality	Style-Adjusted Score
Human Reviewer	1236	1257	1237	1229	1188
Prompt-based	1391	1392	1387	1396	1396
Liang et al.	1237	1251	1242	1217	1227
D'Arcy et al.	1766	1731	1761	1769	1776
MAMORX (Ours)	**1870**	**1869**	**1874**	**1889**	**1914**

Table 2: Comparison of review systems using Elo ratings across different dimensions. Scores are presented as mean $\pm$ standard error of the mean.

We conducted an experiment in which 13 master's and Ph.D. level students produced 140 judgments in total. The participants chose between two randomly selected reviews for a given scientific paper. The articles and the pair of reviewers to judge were selected using the Elo rating system. Reviewers selected whether either review was superior, if they tied, or if both were poor for each of the four aspects. We computed five scores for each system: Technical Quality, Constructiveness, Clarity, Overall Quality, and Style-Adjusted Score. The Style-Adjusted Score is the Elo rating after removing the impact of style differences between reviewers using the Bradley-Terry model.

In the data, MAMORX lost 12% of the matches against the next-best multi-agent framework, with no losses to all other models. Furthermore, our system won in 88% of cases against a human reviewer. Notably, MAMORX is superior in Elo rating for all aspects, and the style-adjusted score. It is interesting that the style-adjusted score is higher than the unadjusted score, compared to Liang et al., which tended to have a similar style (e.g., length). However, D'Arcy et al. and our system were significantly more preferable to the participants, thus making their adjusted score higher. The human reviews had the lowest Elo rating across the board, which is unexpected. Also, Liang et al. has a lower Elo compared to the prompt-based method, which has significantly less complexity.

Using the Elo rating system, we can also compute the estimated win rate of each system. This is shown in Table 3. Here, the system estimates that MAMORX is preferred over the human reviewer 98% of the time (not significantly different from the 88% from the actual pairwise comparisons). Notably, MAMORX is expected to be preferred over D'Arcy et al. 67% of the time.

Model	Human Reviewer	Prompt-based	Liang et al.	D'Arcy et al.	MAMORX (Ours)
Human Reviewer	50%	28%	52%	4%	2%
Prompt-based	72%	50%	74%	11%	6%
Liang et al.	48%	26%	50%	4%	2%
D'Arcy et al.	96%	90%	96%	50%	33%
MAMORX (Ours)	98%	94%	98%	67%	50%

Table 3: Comparison of review systems using Elo ratings for the overall quality judgement in Table 2. The numbers in the table represent the percentage of times each system is preferred over the others in the pairwise comparisons (the system in the row is preferred over the system in the column).

7 Discussion and Conclusion

In this study, we introduced MAMORX, a multi-agent, multi-modal scientific review generation framework with external knowledge. Our system leverages specialized agents, multi-modal inputs, and external knowledge sources to review scientific papers. We evaluated MAMORX against human-generated reviews and other baseline methods using the Elo rating system. Our results show that MAMORX outperforms other systems in terms of technical quality, constructiveness, clarity, and overall quality. These findings suggest that MAMORX is a promising tool for supporting scientific review.

We can speculate why our system is preferred more to humans. Due to the rapid growth in scientific literature, humans can no longer consider all previous literature during the peer review process [23, 22, 33]. This leads to AI generated review systems. With the addition of a literature search system, MAMORX is utilizing all the existing literature to perform the novelty assessment as opposed to relying on the limited knowledge within the LLMs. This leads to more concrete novelty assessment. Additionally, MAMORX produces individualized reviews for each figure which helps in directing the improvement of figures within the paper under review. This is likely the cause of favor toward reviews generated by MAMORX. The incorporation of novelty assessment and figure criticism addresses limitations stated within previous AI review systems [24, 10].

We found some limitations in our AI-review systems. One of them is the reviews are sometimes cutoff mid-sentence. This is caused by the output token limits within the LLMs or the probabilties of the next token to be the stop token. Also, we are aware that there are possible bias within AI review systems towards certain fields and ideas. Acknowleding that humans could also contain such biases [9], this is one possible factor to consider in future studies.

Despite the limitations discussed, our work with MAMORX represents a significant advancement in automated scientific review generation. The system's superior performance across multiple dimensions—technical quality, constructiveness, clarity, and overall quality—show its potential to support the current grow of science. By effectively combining multi-agent architecture, multi-modal inputs, and external knowledge sources, MAMORX addresses many of the shortcomings identified in previous AI review systems [44, 3, 17].

For future work, we plan to investigate the potential reasons why human-generated reviews were consistently given lower scores compared to AI-generated review. This also includes a more diverse range of human evaluators for the reviews. The metrics we might consider are the time spent generating the review, thoroughness of the review and more. Eventually, the result will indicate which concepts or advanced techniques are needed to enhance AI's review capabilities. Implementing fine-tuned models to analyze multi-panel figures and correlate them with textual content could further improve the system's multi-modal capabilities [42, 5]. Information from the author, institution, and funding information could be used to provide a more comprehensive review. This, of course, would not be possible for single-blind reviews. Future research should investigate methods for detecting and mitigating algorithmic bias in review generation, such as implementing fairness-aware machine learning techniques [15]. Additionally, exploring the use of federated learning and anonymization could enable MAMORX to maintain data privacy, potentially allowing collaboration across institutions without compromising sensitive information [45].

Acknowledgements

We thank Mo Zhou for his help during the initial phase of our project and the Science of Science and Computational Discovery Lab at CU Boulder for their help.

References

[1] Olatundun D. Awosanya, Alexander Harris, Amy Creecy, Xian Qiao, Angela J. Toepp, Thomas McCune, Melissa A. Kacena, and Marie V. Ozanne. The utility of ai in writing a scientific review article on the impacts of covid-19 on musculoskeletal health. *Current Osteoporosis Reports*, 22(1):146–151, jan 13 2024.

[2] Seyyed Kazem Banihashem, Nafiseh Taghizadeh Kerman, Omid Noroozi, Jewoong Moon, and Hendrik Drachsler. Feedback sources in essay writing: peer-generated or ai-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), apr 2024.

[3] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. *Your Paper has been Accepted, Rejected, or Whatever: Automatic Generation of Scientific Paper Reviews*, pages 19–28. Springer International Publishing, 2016.

[4] David Carabantes, José L. González-Geraldo, and Gonzalo Jover. Chatgpt could be the reviewer of your next scientific paper. evidence on the limits of ai-assisted academic reviews. *El Profesional de la información*, sep 26 2023.

[5] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A Review of Multi-Modal Large Language and Vision Models. 2024.

[6] Alessandro Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 2021.

[7] J. Chen and H. Zhuge. Extractive summarization of documents with images based on multi-modal rnn. *Future Generation Computer Systems*, 99:186–196, 2019.

[8] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[9] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005, 2015.

[10] Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-Agent Review Generation for Scientific Papers. *arXiv*, 2024.

[11] Regina Dittrich, Reinhold Hatzinger, and Walter Katzenbeisser. Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):511–525, 1998.

[12] C. Duong, T. Nguyen, H. Yin, M. Weidlich, S. Mai, K. Aberer, and Q. Nguyen. Efficient and effective multi-modal queries through heterogeneous network embedding. *Ieee Transactions on Knowledge and Data Engineering*, 34:5307–5320, 2022.

[13] Arpad E. Elo. The uscf rating system. *Chess Life*, XIV(13):2, March 1960. PDF.

[14] Xuemei Gu and Mario Krenn. Interesting scientific idea generation using knowledge graphs and llms: Evaluations with 100 research group leaders, 2024.

[15] Mohammad Hosseini and Serge P. J. M. Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(1), may 2023.

[16] Jingshan Huang and Ming Tan. The role of chatgpt in scientific communication: writing better scientific review articles. *American Journal of Cancer Research*, 2023.

[17] Rahul Jha, Catherine Finegan-Dollak, Ben King, Reed Coke, and Dragomir Radev. Content models for survey generation: A factoid-based evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 441–450. Association for Computational Linguistics, 2015.

[18] X. Jiang, B. Tian, and X. Tian. Retrieval and ranking of combining ontology and content attributes for scientific document. *Entropy*, 24:810, 2022.

[19] Melissa A. Kacena, Lilian I. Plotkin, and Jill C. Fehrenbacher. The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, 22(1):115–121, jan 16 2024.

[20] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[21] Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. Peer review in scientific publications: benefits, critiques, & a survival guide. *Ejifcc*, 25(3):227, 2014.

[22] L. Bornmann, R. Haunschild, and R. Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 2020.

[23] L. Bornmann and R. Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.*, 2014.

[24] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*, 2023.

[25] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830, 10 2023.

[26] Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore, December 2023. Association for Computational Linguistics.

[27] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[28] Y. Long, H. Yu, J. Wang, Y. Ji, and S. Qian. Multi-level multi-modal cross-attention network for fake news detection. *Ieee Access*, 9:132363–132373, 2021.

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

[30] Vini Mehta, Ankita Mathur, A.K. Anjali, and Luca Fiorillo. The application of chatgpt in the peer-reviewing process. *Oral Oncology Reports*, 9:100227, 3 2024.

[31] Anna Nikiforovskaya, Nikolai Kapralov, A. Vlasova, Oleg Shpynov, and A. Shpilman. Automatic generation of reviews of scientific papers. *International Conference on Machine Learning and Applications*, 2020.

[32] Marco Pautasso. Publication growth in biological sub-fields: patterns, predictability and sustainability. *Sustainability*, 4(12):3234–3247, 2012.

[33] R. Perrucci, C. Perrucci, and M. Subramaniam. From Little Science to Big Science. 2017.

[34] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

[35] Myriam Safrai and Kyle E. Orwig. Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by chatgpt-4. *Journal of Assisted Reproduction and Genetics*, 41(7):1871–1880, apr 15 2024.

[36] Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855*, 2023.

[37] P. Wang, S. Li, H. Zhou, J. Tang, and T. Wang. Toc-rwg: explore the combination of topic model and citation information for automatic related work generation. *Ieee Access*, 8:13043–13055, 2020.

[38] Qingyun Wang, Carl Edwards, Heng Ji, and Tom Hope. Towards a human-computer collaborative scientific paper lifecycle: A pilot study and hands-on tutorial. In Roman Klinger, Naozaki Okazaki, Nicoletta Calzolari, and Min-Yen Kan, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 56–67, Torino, Italia, May 2024. ELRA and ICCL.

[39] F. Xie, J. Chen, and K. Chen. Extractive text-image summarization with relation-enhanced graph attention network. *Journal of Intelligent Information Systems*, 61:325–341, 2022.

[40] M. Yan, W. Yu, Q. Shi, and X. Tian. A multimodal retrieval and ranking method for scientific documents based on hfs and xlnet. *Scientific Programming*, 2022:1–11, 2022.

[41] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4v(ision). 2023.

[42] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. 2023.

[43] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 2021.

[44] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, sep 29 2022.

[45] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

A Appendix / supplemental material

A.1 Sytem prompts for specialized agents

Here are the prompts used in our agents. Many are largely inspired by [10].

System Prompt for Review Leader

You are part of a group that needs to perform tasks that involve a scientific paper. You are the review_leader, who's in charge of writing the final reviews. You will need to collaborate with other agents by delegating them tasks involving their expertise and communicate with them.To start your task, you should first draft a high-level plan with a list of steps, concisely describing how you will approach the task. Then, execute the plan. When executing the plan, write the current step you are working on each time you move to the next step to remind yourself where you are. You are allowed to create a sub-plan for a step if it is complicated to do in one pass.

Optionally, it may be helpful to share a plan with other agents to help guide them in some cases. Depending on the task, you may need to do multiple rounds of communications, delegations of tasks and receiving suggestions to make sure you have all the necessary information for the final review; you should arrange follow up delegations or communication to other agents if they provide a bad response or seem to have misunderstood the task. These include asking follow-up questions, clarifying your requests, or engaging in additional discussion to fully reason about the task.

To reduce communication errors, after you send a message you should write a short description of what you expect the response to look like. If the response you get doesn't match your expectation, you should review it and potentially ask follow-up questions to check if any mistakes or miscommunications have occurred. It could be the case that an agent (including yourself) has misread something or made a logic error.
Information about agents: There are {num_agents} agents in the group, including yourself. You are {review_leader}. The other agent(s) are: {clarity_agent, experiments_agent, impact_agent}.

VERY IMPORTANT: Make sure to only draft the high-level plan for the review once at the beginning, avoiding generating duplicate messages.

System Prompt for Clarity Agent

You are part of a group of agents working with a scientific paper. You are highly curious and have incredible attention to detail, and your job is to help ensure that the paper has clearly explained its methods, experimental settings, and key concepts and determine whether the paper is well-organized and can be easily understood and reproduced. You will have access to the full text of the paper using the {paper_read_tool}, but be sure to minimize using that tool to avoid generating redundant tokens. If you merely want to access part of the paper that's relevant to the current discourse, use the {paper_search_tool}. The 'review_leader' will ask questions regarding your feedback concerning the clarity of the paper, make sure to respond it and ask follow-up questions if needed. Scrutinize the paper heavily, identifying any missing details or potential issues that could make it ambiguous or hard to understand. Keep in mind that the issues might not be so obvious in practice, so you should think carefully and explore multiple perspectives and possibilities. In particular, make sure the paper provides all information necessary to implement any proposed methods, including any information on any background concepts needed to understand how the methods work. Also ensure that the paper provides enough information to replicate the experimental settings, including any hyperparameters, equipment and material specifications, or other implementation details. Think of the kinds of questions a scientific paper reviewer might ask, or what they might suggest is confusing or poorly explained in the paper. Always make sure that you understand the terms and concepts used in the paper. If you are unsure about the definition of a term or how it is meant to be interpreted in a particular context, you should ask about it, as it is important for the paper to explain such things. When you are done talking with the review_leader, tell it that you are done with your reviewfeedbacks, and give them a summary list of any missing or misleading information, ambiguous statements, poorly organized points, or other suggestions that you identified."

System Prompt for Impact Agent

You are part of a group of agents working with a scientific paper. You are highly curious and skeptical, focusing on the paper's novelty, significance, and impact. Your job is to ensure the paper clearly explains its motivations, goals, and key findings, and to determine if it makes a significant contribution to its field. You will have access to the full text of the paper using the {paper_read_tool}, but minimize its use to avoid redundant tokens. For specific sections, use the {paper_search_tool}. The 'review_leader' will ask for your feedback on the paper's impact and novelty; respond and ask followup questions if needed. Scrutinize the paper for hidden assumptions or issues that could undermine its claimed goals and motivations. Consider multiple perspectives and explore various possibilities. Think of questions a scientific paper reviewer might ask about confusing or poorly justified aspects. Ensure you understand all terms and concepts used. If unsure about a term's definition or interpretation, ask for clarification, as it's crucial for the paper to explain these clearly. When finished reviewing, inform the review_leader and provide a summary list of any missing information, poorly justified points, or other suggestions you've identified to help compose the final review."

System Prompt for Experiment Agent

You are part of a group of agents that must perform tasks involving a scientific paper. You are an expert scientist that designs high-quality experiments, ablations, methodology and analyses for scientific papers. When the reviewer_leader sends a message to you to ask for assistance in coming up with experiments to include in a paper or judging the quality of experiments or methodology that are in a paper, you should help.

You should ensure that you fully understand the claims and goals of the paper before giving suggestions. You will have access to the full text of the paper using the {paper_read_tool}, but be sure to minimize using that tool to avoid generating redundant tokens. If you merely want to access part of the paper that's relevant to the current discussions, use the {paper_search_tool}.

It is crucial to understand what the paper is attempting to investigate in order to design experiments to support the investigation. Obtain any information you need in order to design good experiments, and ask follow up questions if needed.

Be detailed and specific in the experimental suggestions you give. What should the setup be? What settings or methods should be compared? What metrics or measurement techniques should be used? How should the results be analyzed? Make it clear which specific details are important and why (e.g., particular choices of settings, baselines, metrics, environments, procedures, and so on), and which details are unimportant.

If you are asked to check the quality of an existing experimental procedure, one useful approach is to come up with how you would have conducted the experiments and compare the given approach to that in order to generate potential areas for improvement. If you find a shortcoming, explain the issue clearly: why is the existing experiment misleading or why does it fail to fulfill the goals of the investigation?

When you are done talking with the reviewer_leader, tell it that you are done with your reviews/feedbacks, and give them a list of your final feedbacks.

A.2 Sytem prompts for MAMORX with external knowledge

System Prompt for Leader

You are part of a group that needs to perform tasks that involve a scientific paper. You are the review_leader, who's in charge of writing the final reviews. You will need to collaborate with other agents by delegating them tasks inlvoving their expertise and communicate with them. To start your task, your must first draft a highlevel plan with a list of steps,concisely describing how you will approach the task. Then, execute the plan. When executing the plan, write the current step you are working on each time you move to the next step to remind yourself where you are. Optionally, it may be helpful to share a plan with other agents to help guide them in some cases. You must use the {ask_question} and {delegate_work} functions to collaborate with your co-workers. multiple rounds of communications, delegations of tasks and receiving suggestions to make sure you have all the necessary information for the final review your task, ; you should arrange follow up delegations or communication to other agents if they provide a bad response or seem to have misunderstood the task.These include asking follow-up questions, clarifying your requests, or engaging in additional discussion to fully reason about the task.
To reduce communication errors, after you send a message you should write a short description of what you expect the response to look like. If the response you get doesn't match your expectation, you should review it and potentially ask followup questions to check if any mistakes or miscommunications have occurred. It could be the case that an agent (including yourself) has misread something or made a logic error.Information about agents: There are {num_agents} agents in the group, including yourself. You are {review_leader}. The other agent(s) are: {clarity_agent, experiments_agent, impact_agent}. **VERY IMPORTANT**: You must only generate a highlevel plan at the start

A.3 Prompts for Figure Critic

Prompts for Figure Critic

Given the abstract of an academic paper and captions below, generate a short review on the clarity and consistency between the given image with the provided captions and the abstract. Be as critical as possible. If there are any inconsistencies, please list them out as well. Ignore minor inconsistencies since the images might not be complete. Overall, try to give a balanced view and focus on improvement suggestions.
Abstract: <abstract of paper>,
Captions: <a list of captions associated with figures found in the paper>

Prompts to analyze each figure individually

Generate a short description of the provided image. Also describe the implications conveyed within the image

A.4 Prompts for Novelty Tool

Prompts to Generate Search Phrases

Prompt 1: Given the abstract of an academic paper below, generate a search phrase of less than 10 words to find related papers in the field. Return ONLY this phrase. This phrase should be useful for searching for similar papers in academic databases. Use general terms that reflect domain-specific field knowledge to enable a fruitful search.
Abstract: argument['abstract']

Prompt 2: Given the abstract of an academic paper and a previously generated search phrase, create a new, broader search phrase of less than 10 words. This new phrase should expand the search scope to include related concepts or methodologies not covered by the first phrase. Return ONLY this new phrase.
Abstract: argument['abstract']
Previous search phrase: [SearchPhrases]

Prompt 3: Given an academic paper abstract and two previously generated search phrases, create a final, even broader search phrase of less than 10 words. This phrase should capture the most general concepts related to the paper's field of study, potentially including interdisciplinary connections. The goal is to cast the widest possible net for related research. Return ONLY this new phrase.
Abstract: argument['abstract']
Previous search phrase: [SearchPhrases]

Prompt to Filter Relevant Papers

Assess the relevancy of the following paper to the core paper. Be strict in your assessment and only consider it relevant if it closely relates to the core concept. If the core paper and the paper to assess are the same thing, your assessment is "Irrelevant"
Core Paper:
Title: argument['title']
Abstract: argument['abstract']

Paper to Assess:
Title: [title]
Abstract: [abstract]

Provide your assessment as a single word: "Relevant" or "Irrelevant". Only output the single word with no other text or explanation.

Prompt to Assess Novelty

As a skeptical novelty assessor, compare the following proposed academic paper abstract with an existing paper's abstract. Evaluate whether the new paper presents a significantly novel idea or approach compared to the existing paper. It is paramount that you do not let any non-novel paper slip by and look at the overlap through a critical lens.
New Paper:
Title: argument['title']
Abstract: argument['abstract']

Existing Paper
Title: [title]
Abstract: [abstract]

Please consider:
1. A brief comparison of the key ideas, methods, or findings
2. An assessment of the novelty of the new paper compared to the existing one.
3. A clear decision: Is the new paper sufficiently novel compared to this existing paper? Answer with "Novel" or "Not Novel".

However, in your response, simply provide a decision and a 2-3 sentence justification for your decision.
Format your response as follows:
Decision: [Novel/Not Novel]
Justification: [Your Assessment Here]

Prompt to Summarize Novelty Assessment

Given the following novelty assessment results, please summarize whether the proposed paper is novel or not. If any of the comparisons deem the paper as NOT NOVEL, start the summary with 'NOT NOVEL', followed by an explanation that includes the title of the conflicting paper(s). If the paper is considered NOVEL, start the summary with 'NOVEL', and then provide a brief justification of what makes it novel.

Here are the assessment results: [results]