University of Colorado Boulder

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

Pawin Taechoyotin, Mike Wang, Tong Zeng, Bradley Sides, Daniel Acuna
University of Colorado Boulder

Background

The growing amounts of scientific literatures increase the burden of reviewers. Recent studies have explored AI-generated scientific paper reviews. While models like GPT-4 show promise, researchers hold mixed views on this approach.

Major challenges to AI reviews:

- Factual inaccuracies and outdated information
- Reference fabrication
- Weak context understanding
- Inability to provide personalized, constructive feedback

Paper review processes as done by human reviewers often integrate

- textual analysis,
- visual interpretation
- citation assessment
- external knowledge

Recent advances in multimodal AI models (processing text, images, and graphs) and multi-agent systems with external knowledge access offer promising new approaches to address these limitations.

Methods

Multi-agent framework led by a Leader Agent that coordinates specialized agents

- All agents have access to the full text of the paper
- Each agents have different system prompt-driven task
  - Impact Agent: Evaluates significance and novelty
  - Experiments Agent: Critiques methodology, datasets, and experimental design
  - Clarity Agent: Assesses organization, structure, and presentation

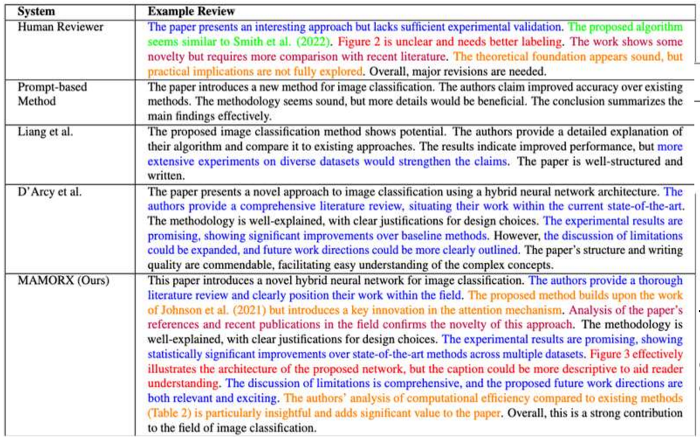
Novel components

- Novelty Assessment: Queries Semantic Scholar to evaluate paper originality
  - Generates queries, builds database of related papers
  - Removes papers already cited by subject paper
  - Performs pairwise novelty assessment against relevant papers
- Figure Critic Assessment: Analyzes visual elements using vision-capable LLMs
  - Extracts figures and captions using PaperMage
  - Evaluates consistency with paper title/abstract
  - Provides clarity assessment and descriptive summaries
- Domain Knowledge: Creates graph of cited works

Technical specification

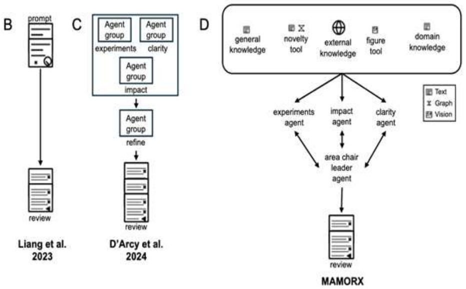
- Model: Claude 3.5 Sonnet
- Agent Framework: CrewAI

Table 1. Comparison of example reviews



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.

Figure 1. Comparison of the architectures of different automated review systems



Evaluation & Results

Evaluation

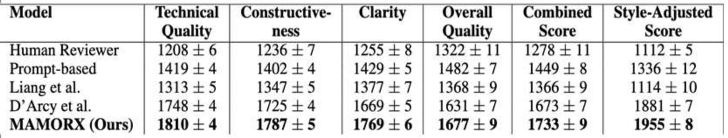
13 graduate students are asked to provide 140 judgments on the reviews

- **Data**: 30 papers (20 with human reviews)
  - 20 from ACL (PeerRead dataset)
  - 10 from NeurIPS 2019
  - One human review randomly selected per paper
- **Elo Rating System**
  - Pairwise comparisons between reviewers
  - Initial rating of 1500 for each reviewer
    - Winner gains points while losers lose points
    - Higher-rated system gain fewer points for beating lower-rated ones
  - Accounting for formatting biases,and covariance adjustment

Results

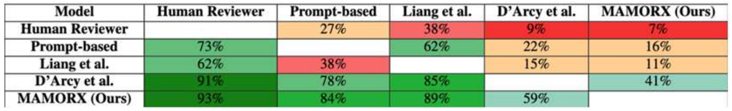
The results are shown in table 2 and table 3. MAMORX achieved the highest combined Elo score showing that multi-agent, multi-modal systems with external knowledge outperform other methods.

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



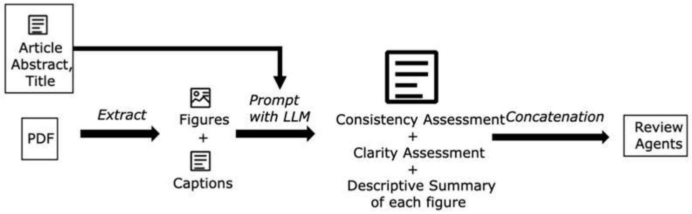
Model	Technical Quality	Constructive-ness	Clarity	Overall Quality	Combined Score	Style-Adjusted Score
Human Reviewer	1208 ± 6	1236 ± 7	1255 ± 8	1322 ± 11	1278 ± 11	1112 ± 5
Prompt-based	1419 ± 4	1402 ± 4	1429 ± 5	1482 ± 7	1449 ± 8	1336 ± 12
Liang et al.	1313 ± 5	1347 ± 5	1377 ± 7	1368 ± 9	1366 ± 9	1114 ± 10
D'Arcy et al.	1748 ± 4	1725 ± 4	1669 ± 5	1631 ± 7	1673 ± 7	1881 ± 7
MAMORX (Ours)	**1810 ± 4**	**1787 ± 5**	**1769 ± 6**	**1677 ± 9**	**1733 ± 9**	**1955 ± 8**

Table 3. Comparison of review systems using Elo ratings for the combined Elo 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)



Model	Human Reviewer	Prompt-based	Liang et al.	D'Arcy et al.	MAMORX (Ours)
Human Reviewer		27%	38%	9%	7%
Prompt-based	73%		62%	22%	16%
Liang et al.	62%	38%		15%	11%
D'Arcy et al.	91%	78%	85%		41%
MAMORX (Ours)	93%	84%	89%	59%	

Figure 2. Comparison of the architectures of different automated review systems
. 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.



Conclusions

MAMORX significantly advances automated scientific review through multi-agent, multi-modal approach with external knowledge integration. Evaluation shows superior performance across all quality metrics compared to human reviewers and previous AI systems

Future work: enhance multi-modal capabilities, and develop bias mitigation techniques