MARS: Multi-Agent Review System for Academic Papers

Kunal Pai[*]
UC Davis
kunpai@ucdavis.edu

Saisha Shetty[*]
UC Davis
spshetty@ucdavis.edu

1 Introduction

Peer review is essential for maintaining the credibility and quality of academic research. However, traditional peer review processes face significant challenges, including reviewer bias, workload imbalances, and inconsistencies in evaluation. With the rapid increase in submissions, the demand for qualified reviewers has outpaced supply, resulting in longer review times and declining quality (McCook, 2006). AI-powered tools, particularly large language models (LLMs) like GPT-3.5 and GPT-4 (Brown et al., 2020; OpenAI, 2024), have been explored as potential solutions, offering efficiency and scalability. Despite their promise, these models have critical limitations that hinder their effectiveness in peer review.

LLMs often produce overly positive and generic feedback while failing to detect methodological flaws, assess novelty, or provide in-depth critiques (Zhou et al., 2024). They struggle with processing long and complex research texts, fact-checking, and identifying related work (Yu et al., 2024; Scherbakov et al., 2024; Ifargan et al., 2025). Furthermore, they lack the domain-specific expertise required to evaluate research soundness and provide meaningful recommendations (Li et al., 2024; Wilkins, 2023).

Given the limitations of single-model LLM reviews, a multi-agent collaboration approach is needed to improve the depth, accuracy, and fairness of AI-assisted peer review. Specialized AI agents can focus on distinct aspects such as methodology evaluation, novelty assessment, and fact verification, mitigating the shortcomings of single-model approaches. Introducing variability in reviewer expertise, tone, and evaluation criteria better mimics real-world peer review dynamics, fostering diverse perspectives.

Additionally, reliance on proprietary LLM APIs raises concerns regarding data privacy, cost, and accessibility for resource-constrained researchers (Finlayson et al., 2024; Staab et al., 2023). A shift towards medium-sized, locally deployable models enhances security and democratizes access to AI-assisted peer review. Structured agent communication and systematic feedback mechanisms further improve review consistency, depth, and transparency, leading to more reliable academic evaluations.

In this work, we present MARS: a Multi-Agent Review System for Academic Papers that leverages a diverse set of specialized AI agents to enhance the peer review process. MARS integrates multiple online and offline models to provide comprehensive feedback on academic submissions, including novelty assessment, grammar checking, and conceptual insights. A multi-agent system comprising reviewer agents with varying expertise, tone, and evaluation criteria ensures robust and nuanced evaluations. These specialized medium parameter agents are aware of each other's decisions and have the agency to accept or reject them, fostering structured and collaborative review cycles.

To enhance the review process, MARS incorporates Socratic questioning, helping reviewers identify weaknesses and gaps in the manuscript. Additionally, it employs a insight fetching model and a novelty model with online capabilities to verify the accuracy of claims and assess contributions against existing literature. To simulate the variability of human reviewers, MARS assigns unique personas to reviewer agents, ensuring diverse perspectives.

The entire pipeline runs on local medium-sized language models, eliminating reliance on external API-based models and enabling resource-constrained researchers to conduct independent peer review.

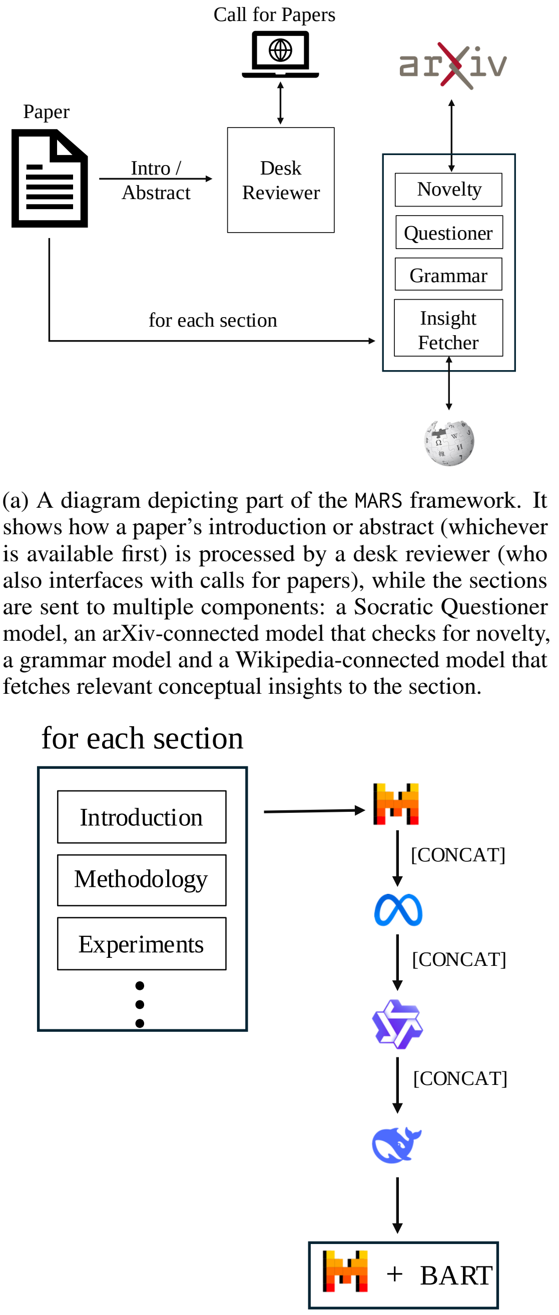
[*] Equal contribution.

2 Related Work

The "data-to-paper" platform employed a multi-agent system where LLM agents handle specific research tasks. Performer agents draft content, reviewer agents refine it, and rule-based agents ensure quality via algorithmic checks like static analysis and runtime error detection. Iterative review cycles integrate LLM and human feedback, breaking down complex tasks into manageable steps. This system enhances research accuracy and quality through automation and verification (Ifargan et al., 2025). While this work focuses on generating sections, we emphasize reviewing existing ones.

(Gottweis et al., 2025) introduced an AI co-scientist for biomedical research that goes beyond literature summarization to generate and refine hypotheses. It uses a multi-agent system for tasks like hypothesis ranking, proximity evaluation, and meta-review, with tools such as web search, domain databases, and AI models like AlphaFold. A natural language interface allows scientists to specify goals and provide feedback, while self-play strategies like scientific debate enhance hypothesis quality. Safety mechanisms ensure responsible research direction. This system contributes to scientific discovery, whereas our work primarily focuses on reviewing written sections.

(Jin et al., 2024) proposed AgentReview, a multi-agent peer review system addressing biases, complexity, and privacy concerns. It follows a structured five-phase pipeline using the OpenAI framework: (I) independent reviewer assessments, (II) author rebuttals, (III) reviewer-AC discussions, (IV) meta-review compilation, and (V) final paper decisions, with a fixed acceptance rate of 32%. While AgentReview explores different reviewer tones, we extend our approach by incorporating online novelty and clarity validation models referencing Wikipedia and arXiv and a dedicated grammar-checking model. Our pipeline supports broader conferences beyond ML/NLP and can be run locally using medium-sized language models.



(a) A diagram depicting part of the MARS framework. It shows how a paper's introduction or abstract (whichever is available first) is processed by a desk reviewer (who also interfaces with calls for papers), while the sections are sent to multiple components: a Socratic Questioner model, an arXiv-connected model that checks for novelty, a grammar model and a Wikipedia-connected model that fetches relevant conceptual insights to the section.

(b) A diagram illustrating the workflow of the section-wise Reviewer language models. Each section is initially reviewed by a model based on Mistral 7B. The review is then appended to the section and passed to the next reviewer, based on Llama 3.2 3B. This process continues sequentially through Qwen2.5 7B and DeepSeek-R1 7B. Finally, a meta-reviewer and a BART summarizer generate a consolidated summary of all the reviews for each section.

Figure 1: MARS Framework

3 Methods

3.1 Framework

MARS is a multi-agent framework to enhance the academic peer review process by incorporating multiple online and offline specialized agents to provide diverse and structured feedback. Agents are specialized through their system message on initialization.

These models can be seen in Figure 1, and information about their data sources and functions can be seen in Table 1. All models chosen are state-of-

Agent	Function	Data Source / Model	Justification
Desk Reviewer	Checks topic relevance and introduction alignment with CFP	CFP URL, Llama 3.2 3B with topics from CFP URL, manuscript intro. / abstract	CFP has all the relevant topics to a conference.
Novelty Reviewer	Identifies research novelty with respect to titles and abstracts of similar papers	arXiv API (5-10 papers) with Llama 3.2 3B	arXiv is a vast repository of recent preprints across multiple scientific disciplines
Grammar Reviewer	Checks grammar and readability.	Llama 3.2 3B-based grammar checker	Grammar is an important aspect of a well-written paper.
Questioner	Generates Socratic questions to challenge assumptions.	Llama 3.2 3B with Socratic prompting	Socratic questioning encourages deeper analysis and alternative perspectives (Chang, 2023).
Insight Fetcher	Fact-checks general claims.	Llama 3.2 3B connected with Wikipedia API.	Provides in-context verification on general terminologies used.
Reviewer Agents	Simulate diverse reviewers with varied tones, expertise, and career stages.	Mistral 7B (Jiang et al., 2024), Llama 3.2 3B (Dubey et al., 2024), Qwen2.5 7B (Yang et al., 2024), DeepSeek-R1 7B (Guo et al., 2024).	Models realistic peer review variability (Jin et al., 2024).
Summarizer	Aggregates and summarizes reviews using sentiment weighting.	Mistral 7B, VADER (Hutto and Gilbert, 2014), BART (Lewis, 2019).	Enhances summary coherence and informativeness (Abbasimehr and Shabani, 2019).

Table 1: Agents and Their Roles in the Review Process

the-art language models known for their superior performance in natural language understanding and generation with their parameter sizes, and enhance the framework by ensuring context-aware and high-quality feedback, critical for nuanced academic evaluations. We will leverage Ollama [1] to run local models for our study.

The questioner model uses Socratic questioning (Elder and Paul, 1998; Paul and Elder, 2007; Katsara and De Witte, 2019) to help reviewer agents determine the quality of the paper. Socratic questioning is known to help with critical thinking and getting to the core of arguments or issues presented in a manuscript. By generating a series of section-wise targeted and structured questions, the questioner model aids reviewers in identifying potential weaknesses, ambiguities, or gaps in the manuscript. These questions also encourage authors to provide more clarity or evidence in their revisions.

To enhance the review pipeline's effectiveness, Wikipedia and arXiv, through the insight fetcher and novelty models serve complementary purposes. While Wikipedia verifies general claims and common knowledge, arXiv identifies cutting-edge contributions and assesses novelty. Wikipedia's broad coverage and frequent updates ensure accuracy for widely accepted facts, while arXiv offers access to recent preprints in scientific domains, making it ideal for identifying overlaps with recent research.

Each reviewer agent is assigned a unique persona, such as an early-career researcher, an experienced senior researcher, or an industry leader, with

[1] https://ollama.com/

tones like supportive, critical, neutral and knowledge levels from novice to expert, to simulate the diversity of opinions typically found in the peer review process. These personas influence their evaluation criteria and feedback style, introducing a degree of randomness and variability that mirrors real-world peer review dynamics. This randomness is informed by studies on the inherent subjectivity of peer review (Jin et al., 2024; Scherbakov et al., 2024).

Multi-Agent Communication: Our pipeline involved a sequential review process, where each model received feedback from the previous reviewers and had the autonomy to agree or disagree while generating its own assessment. This structure fostered collaborative evaluation while preserving independent decision-making. The models followed a fixed order: *Mistral* → *Llama 3.2* → *Qwen2.5* → *DeepSeek-R1*. The workflow can also be seen in Figure 1b.

3.2 Evaluation

Accuracy: To validate our system's performance, we evaluated MARS against a ground-truth dataset of 10 previously reviewed papers from ICLR 2023.

We evaluated against state-of-the-art single-model baselines for paper review. Specifically, we select OpenAI's o3-mini (OpenAI, 2025a,b) for its strong reasoning capabilities and NotebookLM (Google, 2023), which is built on Gemini 1.5 (Team et al., 2024) and designed for multimodal in-context ("grounded") question-answering. NotebookLM leverages Gemini 1.5's superior understanding of multimodal documents (Thelwall, 2024) to provide responses contextualized to an uploaded paper, increasing the likelihood that its review remains closely tied to the document.

For both baseline models, we employed a structured prompt to elicit realistic, binary reviews: "You are a pragmatic reviewer for ICLR 2023. Review each academic paper and provide an accept/reject decision with a detailed justification, focusing solely on the content of the paper."

To compute an accept or reject score for MARS, we implemented the following methodology:

Let r_i be the score for section i and w_i its corresponding weight. The overall weighted score S is defined as:

$$S = \frac{\sum_{i=1}^{n} w_i r_i}{\sum_{i=1}^{n} w_i}.$$

We assigned every section an equal weight of $w_i =$ 1, so the formula simplifies to:

$$S = \frac{1}{n} \sum_{i=1}^{n} r_i.$$

Every instance of "Accept" adds a 1 to the score, while a "Reject" does not add anything.

The final decision D based on S is given by:

$$D = \begin{cases} \text{Accept} & \text{if } S \geq 50, \\ \text{Reject} & \text{if } S < 50. \end{cases}$$

We evaluate the performance of the MARS prediction system by comparing its output to the ground truth decisions.

Review Quality: To assess the quality of MARS reviews with the ground truth human reviews, we explored three widely-used metrics for comparing LLM-generated content with human counterparts: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). BLEU calculates n-gram precision between generated and reference texts, which is useful for evaluating lexical alignment but exhibits limitations with paraphrasing variations. ROUGE-L is effective in capturing structural similarities through longest common subsequence identification, thus accommodating sentence reformulations common in texts. METEOR incorporates stemming and synonym matching, addressing semantic equivalence beyond surface-level matches. While no single metric perfectly captures human judgment of similarity, a combined approach is more robust.

We define a high-quality review as one that should effectively capture the core ideas and reasoning found in human-written reviews without relying on direct memorization of specific phrases. This means that a good review should achieve relatively high METEOR and ROUGE-L scores, as these metrics emphasize semantic similarity and structural alignment. A strong METEOR score suggests that the review preserves meaning through synonym matching and stemming, ensuring that the same concepts are conveyed even if the wording differs. Likewise, a high ROUGE-L score indicates that the generated review shares key structural elements with human reviews, allowing for natural variations in phrasing while maintaining coherence. At the same time, a lower BLEU score can be desirable, as it suggests that the review is not simply replicating exact n-grams from the reference text but instead demonstrating original language generation. This balance ensures that the review remains faithful to the intent and depth of human feedback while avoiding direct plagiarism or excessive rigidity in phrasing. By considering all three metrics together, we can better assess whether a review effectively mirrors human judgment while maintaining originality and meaningful variation.

Since the format of human reviews and LLM-generated reviews did not always match, we first conducted a similarity search to identify the LLM-generated reviews that most closely resembled each human review. Once the closest match was identified, we computed the similarity metrics. This approach helps to ensure that the evaluation reflects how well the LLM-generated reviews align with human reviews without universally penalizing them for formatting or structural differences.

4 Experiments

4.1 Results

Resources: The execution time of the pipeline is primarily influenced by the number of sections in the paper, as each section undergoes independent processing. This section-wise execution introduces a scalability challenge: as the number of sections increases, the total runtime grows proportionally. Additionally, the computational resources available on the system play a significant role in overall performance. Factors such as processor speed, memory availability, and model optimization techniques affect execution time.

For our evaluation, we ran the MARS system on MacBooks equipped with *Apple Silicon*[2], a consumer-grade device accessible to many researchers. Empirical observations indicate that the pipeline takes approximately *8–10 minutes per section*, including model inference and the application of the BART summarizer. This runtime, while scaling with document length, remains manageable even on personal computing devices, further reinforcing the practicality of our approach.

Accuracy: Our Desk Reviewer model consistently returned an “Accept” decision for all submissions when evaluated against ICLR 2023’s Call for Papers, indicating that, according to the specified criteria, all papers were considered relevant to the conference. To further assess its ability to differentiate between suitable and unsuitable submissions, we tested a paper from the dataset against ASPLOS 2025’s Call for Papers [3]. In this case, the model did not classify the paper as “Accept”, demonstrating its capacity to infer relevance across different conferences.

The holistic decision-making performance of the MARS prediction system was evaluated by comparing its output to the ground truth decisions for a set of 10 submitted papers to ICLR 2023, as shown in Table 2. In this experiment, MARS achieved perfect accuracy in classifying both accepted and rejected papers. Specifically, all papers that were accepted in the ground truth had a MARS score greater than or equal to 50, and all rejected papers had a score below 50.

The lowest acceptance score predicted by MARS was 51.94, and the highest was 79.25, demonstrating a clear gradient in the predicted scores for accepted papers. This implies that MARS can not only predict acceptances and rejections but also provide a score of the predicted quality of the submissions, though further analysis is needed to determine the robustness of this score.

While MARS perfectly classified the acceptance and rejection decisions, both NotebookLM and o3-mini exhibited inconsistencies. NotebookLM misclassified three rejected papers as borderline acceptances or weak acceptances, while o3-mini accepted all papers, including those that were rejected in the ground truth. Notably, o3-mini’s tendency to accept all papers raises concerns about its decision boundary, making it less reliable for filtering lower-quality submissions.

Overall, these results show that the MARS prediction system is highly accurate for binary classification, with no misclassifications in this test set. However, to further assess the generalizability of MARS, it would be valuable to test it on a larger, more diverse dataset, including borderline cases and papers with scores near the 50-point threshold. Such an analysis would help to evaluate whether MARS can maintain its accuracy across different types of conferences and submission pools.

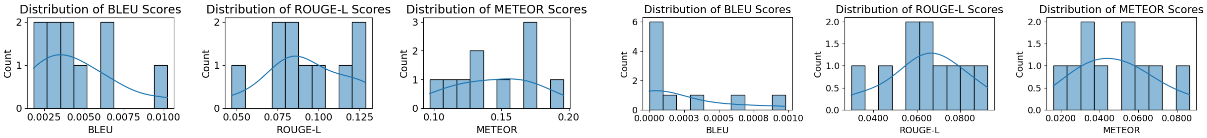
Review Quality: The distributions of BLEU, ROUGE-L, and METEOR scores for LLM-generated reviews compared to ground-truth human reviews are shown in Figure 2. Figure 2a presents the distributions of these scores for concatenated summaries, while Figure 2b displays the distributions for complete reviews.

[2] https://www.apple.com/newsroom/2020/06/apple-announces-mac-transition-to-apple-silicon/

[3] ASPLOS 2025 Call for Papers: https://www.asplos-conference.org/asplos2025/cfp/

Paper	Ground Truth	NotebookLM	o3-mini	MARS
Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections	Accept: poster	Weak Accept	Accept	Accept (Score: 64.00)
GuardHFL: Privacy Guardian for Heterogeneous Federated Learning	Reject	Accept	Accept	Reject (Score: 31.29)
Stochastic No-regret Learning for General Games with Variance Reduction	Accept: poster	Accept (with reservations)	Accept	Accept (Score: 73.00)
Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods	Accept: poster	Accept	Accept	Accept (Score: 79.25)
Exploring perceptual straightness in learned visual representations	Accept: poster	Weak Accept	Accept	Accept (Score: 51.94)
EfficientTTS 2: Variational End-to-End Text-to-Speech Synthesis and Voice Conversion	Reject	Weak Accept	Accept	Reject (Score: 43.06)
Variational Imbalanced Regression	Reject	Borderline Accept	Accept	Reject (Score: 31.91)
Corrupted Image Modeling for Self-Supervised Visual Pre-Training	Accept: notable-top-25%	Accept	Accept	Accept (Score: 57.99)
Communication-Efficient and Drift-Robust Federated Learning via Elastic Net	Reject	Reject	Accept	Reject (Score: 39.20)
FedAvg Converges to Zero Training Loss Linearly: The Power of Overparameterized Multi-Layer Neural Networks	Reject	Reject	Accept	Reject (Score: 26.86)

Table 2: Comparison of ground truth decisions with SoTA single-model decisions and MARS predictions for paper acceptance.



(a) A histogram depiction of the distribution of BLEU, ROUGE-L, and METEOR of the section-wise review summaries of MARS with the human reviews.

(b) A histogram depiction of the distribution of BLEU, ROUGE-L, and METEOR of the complete section-wise reviews of MARS with the human reviews.

Figure 2: MARS Review Quality

Overall, the scores indicate a low to moderate alignment between LLM-generated content and human-written reviews, but they are not particularly high. One reason for this is that MARS' reviews tend to be longer and more comprehensive, making it more challenging for the models to achieve high similarity scores with a more concise human review. Among the three metrics, METEOR achieves the best performance, likely due to its incorporation of synonym matching and stemming, which makes it more robust to variations in word choice. Notably, concatenated summaries yield higher scores than complete reviews, suggesting that summarization improves alignment by capturing key information across all sections in a more concise format. Additionally, the low BLEU scores suggest that the models are not merely memorizing and regurgitating content, but rather are generating responses dynamically.

Looking at the reviews from a human perspective provides valuable insights into how MARS arrives at its conclusions. By prompting the LLM to justify its reasoning based on its expertise and tone, and either agree or disagree with previous decisions, enhances explainability, as it reveals the chain of thought leading to the final review rather than presenting an opaque verdict.

While evaluating MARS reviews, we observed that although the generated critiques were generally well-articulated and structured, they occasionally lacked depth in technical analysis, particularly when assessing novel contributions. However, this limitation is mitigated by the novelty and insight fetcher models. This observation aligns with prior findings that LLMs struggle with abstract reasoning and domain-specific expertise beyond their training data (Xiong et al., 2024). Nonetheless, when submissions followed well-established patterns or addressed common research themes, MARS provided insightful feedback on par with human reviewers.

In summary, while MARS demonstrates a reasonable level of alignment with human reviews based on similarity metrics, its true strength lies in its structured reasoning and consistency. The Socratic questioning component, although not directly influencing final decisions, plays a crucial role in improving the interpretability of the generated reviews by acting as subjective feedback on questions researchers could address. Future iterations of MARS could benefit from integrating domain-specific fine-tuning and additional layers of human oversight to further improve review quality and explainability.

4.2 Multi-Agent Communication

Table 3 presents the probability distribution of each model's decisions.

Mistral's decision serves as the branching point in our decision tree. When it *Rejects* a submission, subsequent reviewers mostly tend to follow this stance. Conversely, when Mistral *Accepts*, the downstream decisions exhibit greater variability.

Llama 3.2 demonstrates a strong tendency to reinforce the initial decision. For example, when Mistral rejects, Llama 3.2 follows suit with a *0.96 probability (24/25 cases)*. Even in the Accept branch, it aligns with Mistral's choice in *74% of cases*. However, there is a chance that it can switch its decision as well.

Qwen2.5's behavior is particularly notable. It exhibits a high degree of conformity to both Mistral and Llama 3.2. When in the *Reject* branch, it rejects with a *0.88 probability*, and in the *Accept* branch, it accepts *81% of the time*. However, a deeper analysis reveals that Qwen2.5 is most strongly influenced by the *most recent* reviewer. When Mistral *Accepts* but Llama 3.2 *Rejects*, Qwen2.5 also *Rejects with 88% probability*, highlighting its tendency to defer to Llama 3.2.

As the final reviewer, DeepSeek-R1 demonstrates a bias toward *acceptance*, even in cases where prior models lean toward rejection. For instance, in one *Reject* branch, it still *accepts with 69% probability*, occasionally reversing earlier decisions. When the process starts with an *Accept* from Mistral, DeepSeek-R1's *acceptance rate increases further (0.81–1.00)*. This suggests a propensity toward *overly positive feedback*, rather than a more balanced, pragmatic review approach that is expected from a rigorous review process of an academic paper.

4.3 Ablation Studies

Effect of Threshold Values: We evaluated different threshold settings to determine their impact on classification accuracy. With a high threshold (scores above 70 classified as "Accept"; otherwise "Reject"), the model achieved 70% accuracy. Conversely, using a low threshold (scores above 30 classified as "Accept") resulted in 60% accuracy. Given these outcomes, our chosen threshold of 50% strikes a balance and is validated as an appropriate choice. A detailed results table is provided in Appendix A.

Effect of Reviewer Models and Non-Reviewer Models: We evaluated the performance of different model groups: (1) the reviewer models alone, (2) all other models without the reviewer (i.e., the insight fetcher, novelty model, grammar model, and summarizer), (3) the insight fetcher and novelty models, (4) only the insight fetcher model, (5) only the novelty model, (6) only the grammar model, and (7) only the summarizer model.

Our findings reveal that only the reviewer models achieve fully accurate predictions. In contrast, the non-reviewer models perform significantly worse, correctly classifying papers only 50% of the time, and they exhibit a strong bias toward predicting "Accept" for all submissions. While the insight fetcher and novelty models can generate both "Accept" and Reject" labels, their accuracy remains limited at around 60%. The insight fetcher model alone achieves only 50% accuracy due to its consistent Reject" predictions, while the novelty model, capable of binary decisions, still only reaches 60% accuracy.

The grammar and summarizer models further demonstrate the limitations of single-purpose models in making robust acceptance decisions. The grammar model achieves 50% accuracy but exhibits a strong bias, consistently predicting "Accept" for all submissions, making it unreliable as a standalone reviewer. The summarizer model performs slightly better, with 60% accuracy, but its

mistral	**llama3.2**	**qwen2.5**	**deepseek-r1**
Reject	Reject (0.96, 24/25)	Reject (0.88, 21/24)	Accept: 0.69 (11/16) Reject: 0.31 (5/16)
Reject	Reject (0.96, 24/25)	Accept (0.12, 3/24)	Accept: 1.00 (1/1)
Reject	Accept (0.04, 1/25)	Reject (1.00, 1/1)	Accept: 1.00 (1/1)
Accept	Reject (0.26, 25/96)	Reject (0.88, 22/25)	Accept: 0.81 (13/16) Reject: 0.19 (3/16)
Accept	Reject (0.26, 25/96)	Accept (0.12, 3/25)	Accept: 1.00 (2/2)
Accept	Accept (0.74, 71/96)	Accept (0.81, 57/70)	Accept: 0.86 (31/36) Reject: 0.14 (5/36)
Accept	Accept (0.74, 71/96)	Reject (0.19, 13/70)	Accept: 0.88 (7/8) Reject: 0.12 (1/8)

Table 3: Conditional probabilities for reviewers.
Rows are branches of the decision tree: the first column indicates the value of *mistral*, while the subsequent columns list the outcome (and probability with counts) for *llama3.2*, *qwen2.5*, and *deepseek-r1* respectively.

decision-making is heavily skewed. It accepts all papers except for one that received a unanimous rejection from all other models, indicating that it only rejects submissions with overwhelmingly negative content.

The poor performance of non-reviewer models suggests that factors such as factual accuracy, novelty, grammar, and summarization quality, while important, are insufficient on their own to make reliable acceptance decisions. Relying on arXiv to identify similar papers is a reasonable approach but may benefit from more advanced retrieval methods. Additionally, the insight fetcher model requires significant improvements, as relying on just a few sources from Wikipedia to contextualize a paper within a broader knowledge base is not optimal. The grammar and summarizer models, despite their complementary roles, lack the nuanced reasoning capabilities necessary for peer review. This underscores the complexity of the review process, where human-like judgment, likely captured better by the reviewer models, is essential for well-calibrated assessments. A detailed results table is provided in Appendix B.

Effect of Reviewer Language Models: We evaluated the performance of our reviewer language models: *mistral*, *llama3.2*, *qwen2.5*, and *deepseek-r1*. *llama3.2* and *qwen2.5* achieve 100% accuracy, meaning they correctly interpret the context set by preceding models (*mistral* for *llama3.2* and both *mistral* and *llama3.2* for *qwen2.5*), aligning or disagreeing with prior assessments as appropriate. *deepseek-r1* exhibits a strong bias toward classifying most papers as "Accept," leading to an accuracy of only 50%. This suggests that despite its superior reasoning capabilities or potential compression into a smaller parameter model, it may lack the nuanced critique necessary for rigorous academic paper reviews. *mistral* tends to be more critical, accepting only 40% of papers. However, its overall accuracy remains at 50%, indicating that its stricter evaluation does not necessarily translate into better judgment. A detailed results table can be found in Appendix C.

5 Conclusion

In this work, we introduced MARS, a multi-agent review system that uses large language models to simulate the peer review process. By incorporating specialized models for question generation, conceptual insight retrieval, and novelty assessment, MARS provides comprehensive evaluations that mirror human reviewer dynamics. Our system demonstrated high accuracy in predicting paper acceptance decisions, highlighting its potential to streamline the review process and reduce reviewer workload.

References

Hossein Abbasimehr and Mostafa Shabani. 2019. A sentiment aggregation system based on an owa operator. In *2019 5th International Conference on Web Research (ICWR)*, pages 1–5.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Edward Y Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0351–0360. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Linda Elder and Richard Paul. 1998. The role of socratic questioning in thinking, teaching, and learning. *The Clearing House*, 71(5):297–301.

Matthew Finlayson et al. 2024. Logits of api-protected llms leak proprietary information. *arXiv preprint arXiv:2403.09539*.

Google. 2023. Introducing notebooklm. Google Blog.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2025. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1):AIoa2400555.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. *Preprint*, arXiv:2406.12708.

Ourania Katsara and Kristof De Witte. 2019. How to use socratic questioning in order to promote adults' self-directed learning. *Studies in the Education of Adults*, 51(1):109–129.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Michael Li, Jianping Sun, and Xianming Tan. 2024. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Systematic Reviews*, 13(1):219.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alison McCook. 2006. Is peer review broken? submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. what's wrong with peer review? *The scientist*, 20(2):26–35.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025a. Openai o3-mini: Pushing the frontier of cost-effective reasoning.

OpenAI. 2025b. Openai o3-mini system card. Technical report, OpenAI.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Richard Paul and Linda Elder. 2007. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36.

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. 2024. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. *arXiv preprint arXiv:2409.04600*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

M Thelwall. 2024. Is google gemini better than chatgpt at evaluating research quality? *Journal of Data and Information Science*.

David Wilkins. 2023. Automated title and abstract screening for scoping reviews using the gpt-4 large language model. *arXiv preprint arXiv:2311.07918*.

Kai Xiong, Xiao Ding, Ting Liu, Bing Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Yixin Cao. 2024. Meaningful learning: Enhancing abstract reasoning in large language models via generic fact guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. 2024. Is your paper being reviewed by an llm? investigating ai text detectability in peer review. *arXiv preprint arXiv:2410.03019*.

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.

A Threshold Evaluation

In this section, we provide the data for testing various thresholds. For a high threshold, any score above a 70 is considered an "Accept" ("Reject" otherwise), and for a low threshold any score above a 30 is considered an "Accept" ("Reject" otherwise). The results are in Table 4.

B Reviewer Models vs. Non-Reviewer Models

In this section, we provide the data for testing only the reviewer models and only the non-reviewer models, namely the grammar, novelty and insight fetcher models. The results are in Table 5.

C Reviewer Language Models

In this section, we provide the data for testing the accuracy of the language models used as the reviewers. The results are in Table 6.

Paper	**Higher Threshold**	**Lower Threshold**
Corrupted Image Modeling for Self-Supervised Visual Pre-Training	Reject	**Accept**
Stochastic No-regret Learning for General Games with Variance Reduction	**Accept**	**Accept**
Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections	Reject	**Accept**
EfficientTTS 2: Variational End-to-End Text-to-Speech Synthesis and Voice Conversion	**Reject**	Accept
Exploring perceptual straightness in learned visual representations	Reject	**Accept**
GuardHFL: Privacy Guardian for Heterogeneous Federated Learning	**Reject**	Accept
FedAvg Converges to Zero Training Loss Linearly: The Power of Overparameterized Multi-Layer Neural Networks	**Reject**	**Reject**
Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods	**Accept**	**Accept**
Variational Imbalanced Regression	**Reject**	Accept
Communication-Efficient and Drift-Robust Federated Learning via Elastic Net	**Reject**	Accept

Table 4: Threshold Evaluations. Boldface represents predictions that were correct

Paper	**Reviewer Models**	**Non-Reviewer Models**	**Insight Fetcher and Novelty Models**	**Only Insight Fetcher Model**	**Only Novelty Model**	**Only Grammar Check**	**Only Summarizer**
Corrupted Image Modeling for Self-Supervised Visual Pre-Training	**Accept**	**Accept**	Reject	Reject	Reject	**Accept**	**Accept**
Stochastic No-regret Learning for General Games with Variance Reduction	**Accept**	**Accept**	**Accept**	Reject	**Accept**	**Accept**	**Accept**
Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections	**Accept**	**Accept**	**Accept**	Reject	**Accept**	**Accept**	**Accept**
EfficientTTS 2: Variational End-to-End Text-to-Speech Synthesis and Voice Conversion	**Reject**	Accept	Accept	**Reject**	Accept	Accept	Accept
Exploring perceptual straightness in learned visual representations	**Accept**	**Accept**	Reject	Reject	Reject	**Accept**	**Accept**
GuardHFL: Privacy Guardian for Heterogeneous Federated Learning	**Reject**	Accept	Accept	**Reject**	Accept	Accept	Accept
FedAvg Converges to Zero Training Loss Linearly: The Power of Overparameterized Multi-Layer Neural Networks	**Reject**	Accept	**Reject**	**Reject**	**Reject**	Accept	**Reject**
Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods	**Accept**	**Accept**	**Accept**	Reject	**Accept**	**Accept**	**Accept**
Variational Imbalanced Regression	**Reject**	Accept	**Reject**	**Reject**	**Reject**	Accept	Accept
Communication-Efficient and Drift-Robust Federated Learning via Elastic Net	**Reject**	Accept	**Reject**	**Reject**	**Reject**	Accept	Accept

Table 5: Reviewer and Non-Reviewer Model Evaluations. Boldface represents predictions that were correct.

Paper	**Mistral**	**Llama3.2**	**Qwen2.5**	**DeepSeek-R1**
Corrupted Image Modeling for Self-Supervised Visual Pre-Training	Reject	**Accept**	**Accept**	**Accept**
Stochastic No-regret Learning for General Games with Variance Reduction	**Accept**	**Accept**	**Accept**	Reject
Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections	Reject	**Accept**	**Accept**	**Accept**
EfficientTTS 2: Variational End-to-End Text-to-Speech Synthesis and Voice Conversion	Accept	**Reject**	**Reject**	Accept
Exploring perceptual straightness in learned visual representations	Reject	**Accept**	**Accept**	**Accept**
GuardHFL: Privacy Guardian for Heterogeneous Federated Learning	**Reject**	**Reject**	**Reject**	Accept
FedAvg Converges to Zero Training Loss Linearly: The Power of Overparameterized Multi-Layer Neural Networks	**Reject**	**Reject**	**Reject**	Accept
Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods	**Accept**	**Accept**	**Accept**	**Accept**
Variational Imbalanced Regression	Accept	**Reject**	**Reject**	**Reject**
Communication-Efficient and Drift-Robust Federated Learning via Elastic Net	**Reject**	**Reject**	**Reject**	Accept

Table 6: Reviewer Language Model Evaluations. Boldface represents predictions that were correct