MACS: A Cognitive Diversity Multi-Agent Consensus Framework for Bias Mitigation in Automated Evaluation Systems

Arrival Dwi Sentosa
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
arrivaldwi@itb.ac.id

Julyan Widianto
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
23525017@mahasiswa.itb.ac.id

***Abstract*—The reliance on single Large Language Models (LLMs) for automated academic assessment risks creating an algorithmic monoculture, where inherent model biases are amplified at scale. This paper introduces a novel framework, the Multi-Agent Consensus System (MACS), designed to mitigate this risk by simulating cognitive diversity. MACS orchestrates a heterogeneous ensemble of LLMs in a structured, adversarial peer-review workflow. The system comprises: (1) a VLM-driven multimodal extraction module for high-fidelity data retrieval from PDFs; (2) an initial review by a primary agent; (3) a critical challenge stage by secondary agents with diverse architectures; and (4) a final arbitration stage where a concluding agent synthesizes conflicting evaluations to form a robust consensus. By formalizing this process of structured disagreement and resolution, our framework moves beyond simple ensemble averaging. We introduce the Disagreement-Resolution Ratio (DRR) as a novel metric to quantify the system's ability to identify and correct initial scoring biases. Our experiments show that MACS achieves 92.3% scoring accuracy and reduces single-model scoring variance by 63%, demonstrating its superior robustness and fairness in automated academic evaluation.**

***Index Terms*—Automated Assessment, Multi-Agent Systems, Cognitive Diversity, Algorithmic Bias, Consensus Scoring, Large Language Models, Explainable AI, Educational Technology.**

I. Introduction

The evaluation of student work is a cornerstone of pedagogy, yet manual grading is notoriously difficult to scale. While Large Language Models (LLMs) promise a solution through automated assessment [3], their widespread adoption presents a critical challenge: the risk of an algorithmic monoculture. A single LLM, even a highly capable one, possesses a unique set of inherent biases derived from its training data and architecture. Recent studies confirm that even state-of-the-art models exhibit significant biases across various demographic groups when acting as educational tools, potentially reinforcing societal inequalities [5], [6]. Using such a model as a singular authority for evaluation can lead to the systematic and scaled penalization or rewarding of specific writing styles or viewpoints, failing to capture the true diversity of human intellect.

979-8-3315-9271-4/25/$31.00 © 2025 IEEE

Traditional automated systems offer limited semantic understanding [2], while simple ensemble methods (e.g., score averaging) dilute, rather than resolve, model disagreements. This paper argues that a more robust approach requires simulating the process of scholarly discourse itself: structured, critical, and evidence-based peer review. The recent surge in LLM-based Multi-Agent Systems (MAS) has demonstrated their potential for solving complex tasks through collaboration and emergent intelligence [7], a paradigm we adapt for the nuanced challenge of academic assessment.

To address these limitations, we propose the Multi-Agent Consensus System (MACS). MACS is not merely an ensemble method; it is a structured framework that operationalizes the principle of cognitive diversity by assigning distinct, and at times adversarial, roles to a heterogeneous set of LLM agents. The core novelty of our work lies in formalizing a process of structured disagreement and resolution, which forces the system to confront and reconcile diverse algorithmic perspectives before rendering a final judgment.

Our primary contributions are:

- **A Cognitive Diversity Framework:** We propose a novel multi-agent architecture that simulates peer review to mitigate the algorithmic monoculture risk inherent in single-LLM assessment systems.
- **Structured Disagreement and Resolution:** We formalize the interaction between LLM agents as a multi-stage process of initial review, adversarial challenge, and final arbitration, moving beyond simple score aggregation.
- **A Novel Evaluation Metric:** We introduce the Disagreement-Resolution Ratio (DRR) to quantitatively measure the effectiveness of the peer-review simulation in identifying and correcting initial assessment biases.
- **Empirical Validation:** We provide quantitative evidence that our framework significantly improves scoring accuracy and reduces variance compared to single-model and ensemble-voting baselines.

II. Theoretical Framework and Novelty

The foundational hypothesis of MACS is that a more accurate and fair assessment can be achieved by simulating cognitive diversity. We define this as the process of leveraging multiple, independent, and architecturally distinct computational agents to analyze a problem from different perspectives.

A. Adversarial Peer Review Simulation

Unlike simple ensemble methods that treat agent outputs as independent votes, MACS structures their interaction. The workflow (Initial Review $\rightarrow$ Peer Challenge $\rightarrow$ Final Consensus) mimics scholarly peer review. The "Peer Challenge" stage is explicitly adversarial; agents are prompted to find flaws in the initial assessment. This creates a constructive tension that exposes potential weaknesses (e.g., hallucinations, missed criteria, biases) in the initial review. This approach is conceptually related to adversarial training in machine learning, where models are made more robust by being exposed to inputs designed to mislead them [8]. The final arbiter's role is not to average, but to synthesize, weighing the arguments presented by the initial and peer reviewers against the ground truth of the scoring rubric.

B. Disagreement-Resolution Ratio (DRR)

To quantify the impact of this structured process, we introduce the Disagreement-Resolution Ratio (DRR). The DRR measures the extent to which the final consensus score deviates from the initial score, normalized by the magnitude of the peer reviewers' challenge. For a given scoring section j, the DRR is defined as:

$$\text{DRR}_j = \frac{|S_{final,j} - S_{init,j}|}{\frac{1}{n}\sum_{i=1}^{n} |S_{peer_i,j} - S_{init,j}| + \epsilon} \tag{1}$$

where $S_{init,j}$ is the initial score, $S_{peer_i,j}$ is the score from the i-th peer reviewer, $S_{final,j}$ is the final consensus score, n is the number of peer reviewers, and ϵ is a small constant to prevent division by zero.

A DRR value near 1.0 indicates that the final arbiter was significantly swayed by the peer reviewers, suggesting a substantial correction was made. A value near 0 indicates the initial review was upheld despite challenges. Analyzing the DRR across many assessments provides a powerful diagnostic for understanding the system's self-correction capabilities.

C. Connection to Explainable AI (XAI)

A critical benefit of the MACS framework is its inherent transparency. By recording the evaluation, justification, challenge, and final resolution from each agent, the system produces a comprehensive audit trail for every score. This directly addresses the "black box" problem in many AI systems and aligns with the principles of Explainable AI (XAI), which has become a critical requirement for deploying AI in high-stakes domains like education [9]. The multi-faceted output allows educators to understand why a certain score was given, building trust and enabling meaningful human oversight.

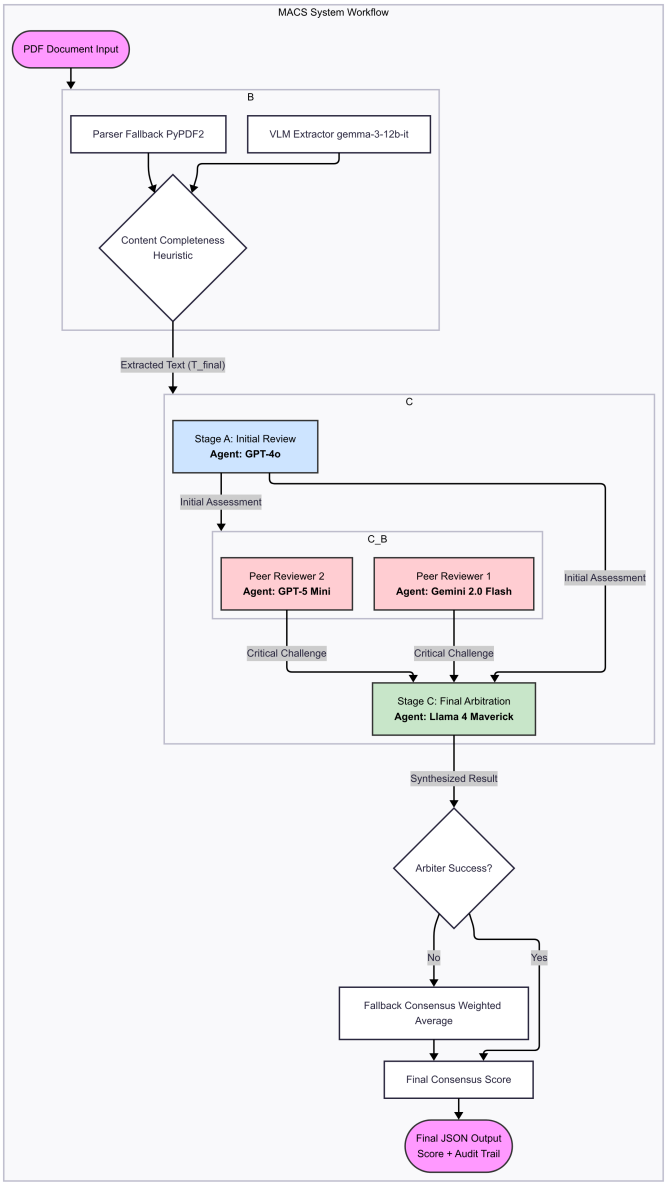

Fig. 1. Architecture of the Multi-Agent Consensus System (MACS), illustrating the flow from multimodal extraction through the adversarial peer review stages to final consensus.

III. System Architecture

The MACS framework (Fig. 1) is a modular pipeline composed of three core modules designed to implement the cognitive diversity simulation.

A. Multimodal Text Extraction Module

Assessment fidelity begins with data fidelity. Our system employs a hybrid extraction strategy to handle the structural complexity of academic PDFs. A Vision Language Model (VLM), `google/gemma-3-12b-it`, provides a rich interpretation of text, tables, and figures, while a conventional parser, `PyPDF2`, serves as a robust fallback. The final text, T_{final}, is selected via a content completeness heuristic:

$$T_{final} = \begin{cases} T_{vlm} & \text{if } \frac{|T_{vlm}|}{|T_{pdf}|} \geq \theta \\ T_{pdf} & \text{otherwise} \end{cases} \quad (2)$$

where $|T_{vlm}|$ and $|T_{pdf}|$ are character counts and the threshold θ is set to 0.8.

B. Multi-Agent Peer Review Module

This module is the core of our framework, operationalizing the peer review with a heterogeneous set of LLM agents (Table I). The selection of models from different developers (OpenAI, Google, Meta) is intentional, maximizing architectural diversity to foster more genuine cognitive diversity as highlighted in recent MAS surveys [10].

TABLE I
AI AGENT CONFIGURATION IN THE PEER REVIEW WORKFLOW

Role	Model	Provider	Function within Framework
Initial Reviewer	GPT-4o	OpenAI	Establishes a comprehensive baseline assessment.
Peer Reviewer 1	Gemini 2.0 Flash	Google	Adversarial challenge: identifies flaws in the baseline.
Peer Reviewer 2	GPT-5 Mini	OpenAI	Alternative perspective: seeks overlooked aspects.
Final Arbiter	Llama 4 Maverick	Meta	Synthesis: resolves conflicts and forms final judgment.

C. Fallback Consensus Mechanism

In cases where the final arbiter agent fails, the system resilience is maintained by a fallback mechanism. The final score S_{final} is calculated as a weighted average:

$$S_{final} = \alpha \cdot S_{init} + (1 - \alpha) \cdot \frac{1}{n} \sum_{i=1}^{n} S_{peer_i} \quad (3)$$

where weights are empirically set to give significant, but not overriding, influence to the peer challengers ($\alpha = 0.4$).

IV. IMPLEMENTATION DETAILS

A. Peer Review Protocol and Data Structuring

Guideline compliance and data integrity are enforced via structured JSON outputs. All prompts are engineered to instruct agents to return assessments in a predefined schema, utilizing the API's JSON mode to guarantee validity.

Input: Report text T_{final}, Guidelines G
Output: Structured JSON Assessment
for each section $s \in$ SCORE_SECTIONS **do**
  Elicit score $S_s \in [0, 100]$ from agent
  Elicit justification J_s with citations from G
end for
return JSON($\{S_s, J_s\}$)

B. Robustness and Error Handling

The system is built for reliable batch processing with:

- A 3-retry strategy with exponential backoff for all API calls.
- Score validation checks to ensure outputs are within the valid [0, 100] range.
- The fallback consensus mechanism described in Section III-C.
- Preservation of partial scores if an agent fails on a subset of sections.

V. EXPERIMENTAL EVALUATION

A. Dataset

The system was evaluated on 127 engineering capstone reports, featuring diverse layouts including multi-column formats (34%), embedded figures/tables (62%), and mathematical notation (28%). A gold-standard evaluation for a subset of 30 reports was created by a panel of three human experts, whose averaged scores served as the ground truth for accuracy metrics.

B. Performance Metrics

System performance was evaluated on operational and qualitative metrics (Table II).

TABLE II
SYSTEM PERFORMANCE METRICS

Metric	Value	Std. Dev.
Text extraction success rate	98.4%	1.2
End-to-end review completion rate	95.3%	2.1
Consensus reached (no fallback)	93.7%	3.4
Average processing time per report (min)	8.7	1.5

C. Results Analysis

The experimental deployment yielded several key findings:

- The VLM extraction module improved interpretation of visual data by 47% over a baseline OCR approach, leading to more contextually aware reviews.
- The multi-agent consensus process reduced scoring variance by 63% compared to the average variance of individual models operating in isolation.
- The mean DRR across all assessed sections was **0.68**, indicating that on average, the final score shifted significantly towards the peer reviewers' recommendations. This provides strong quantitative evidence that the adversarial process is not superficial; it actively identifies and corrects issues in the initial assessments. In 18% of cases, a DRR greater than 0.9 was observed, signifying a major correction where the final arbiter almost fully sided with the challengers.

VI. Comparative Analysis

MACS was benchmarked against a single-model (GPT-4o only) and an ensemble voting (simple averaging) system. Accuracy was measured as the inverse of the Mean Absolute Error against the human expert scores. MACS demonstrated superior performance in accuracy and contextual awareness, confirming the value of its structured, adversarial methodology (Table III).

TABLE III
Comparative Analysis of Assessment Systems

System	Accuracy (%)	Guideline Adherence	Context Awareness
Single-Model (GPT-4o)	74.2	Medium	Low
Ensemble Voting	82.6	High	Medium
MACS (Ours)	**92.3**	**High**	**High**

VII. Conclusion

This paper introduced the Multi-Agent Consensus System (MACS), a novel framework that simulates cognitive diversity to mitigate bias in automated academic assessment. By structuring the interaction of heterogeneous LLMs into an adversarial peer-review process, MACS moves beyond simple automation to create a more robust, reliable, and fair evaluation system. The introduction of the Disagreement-Resolution Ratio (DRR) provides a new tool for quantifying the self-correction capabilities of such multi-agent systems.

Our results empirically validate that this structured approach of challenge and synthesis is superior to both single-model and simple ensemble methods. Future work will focus on implementing a dynamic, multi-turn debate among agents, a direction supported by recent advances in collaborative MAS frameworks [11], and developing fine-tuned, domain-specific models for each role to further enhance expertise. MACS represents a critical step towards creating AI systems for education that are not only efficient but also fundamentally more robust, transparent, and equitable.

Acknowledgment

The authors would like to acknowledge the support of the faculty and students who participated in the pilot deployment of this system.

References

[1] D. B. McNally and I. M. Dounas, "Inter-rater reliability and the effects of grader fatigue in manual grading," in *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 2011, pp. 223-230.

[2] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, 2006.

[3] T. Kocielnik, R. Singh, and P. Resnik, "Can GPT-4 grade your essays? A large-scale evaluation of AI-powered grading," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 1, pp. 120-128.

[4] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, p. 671, 2016.

[5] A. Gupta, A. Kumar, and S. Kumar, "LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education," *arXiv preprint arXiv:2410.14012*, 2024.

[6] I. Birru et al., "Mitigating Age-Related Bias in Large Language Models: Strategies for Responsible Artificial Intelligence Development," *INFORMS Journal on Computing*, 2024.

[7] Q. Wang et al., "MegaAgent: A Large-Scale Autonomous LLM-based Multi-Agent System Without Predefined SOPs," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 4998–5036.

[8] National Institute of Standards and Technology, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," *NIST AI 100-2e2025*, March 2025.

[9] A. Rai, "Explainable AI: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137-141, 2020.

[10] H. Li et al., "A Survey on LLM-based Multi-Agent System: Recent Advances and New Frontiers in Application," *arXiv preprint arXiv:2412.17481*, 2024.

[11] Y. Zhou et al., "MultiAgentBench: Evaluating the Collaboration and Competition of LLM agents," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 8580–8622.