Multi-Agent Systems for Learning Assessment in Education: A Comprehensive Survey

Ruobing Zhang
Beijing Institute Of Petrochemical Technology
Beijing, China
2023540062@bipt.cn

Abstract

This paper presents a systematic survey of Multi-Agent Systems (MAS) for learning assessment, a paradigm addressing the need for personalized, adaptive evaluation at scale. We chart the field's trajectory from foundational, rule-based architectures to dynamic, LLM-integrated ecosystems. By critically analyzing representative systems, we propose a classification based on an evolutionary axis, highlighting the key success factors and inherent limitations of each stage. Our analysis reveals personalized formative assessment as the primary application, with emerging uses in collaborative contexts. However, widespread adoption is hindered by persistent challenges in scalability, interpretability, data privacy, and the risks of algorithmic bias and hallucination. We conclude that MAS are best positioned to augment, not replace, human educators. Finally, we propose a unified research agenda focused on developing the next generation of explainable, ethical, and pedagogically sound assessment tools.

CCS Concepts

• **Applied computing** → Education; Computer-assisted instruction.

Keywords

Multi-Agent Systems (MAS), Large Language Models (LLMs), Learning Assessment, Personalized Education

ACM Reference Format:
Ruobing Zhang. 2025. Multi-Agent Systems for Learning Assessment in Education: A Comprehensive Survey. In *2025 3rd International Conference on Educational Knowledge and Informatization (EKI 2025), July 11–13, 2025, Chongqing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3765325.3765390

1 Introduction

The shift towards personalized learning has created a tension with the logistical challenges of scaling high-quality, individualized assessment. Traditional educational technologies, including conventional Large Language Models (LLMs) reliant on static data, often lack the adaptive and interactive capabilities for meaningful, continuous evaluation [1][2]. This deficit highlights a critical need for intelligent systems that can model learner cognition, provide context-aware feedback, and dynamically adapt educational pathways.

Multi-Agent Systems (MAS)—computational systems composed of multiple autonomous, interacting agents [3]—have emerged as a compelling framework to address this challenge. By simulating complex interactions, MAS can create dynamic and responsive environments for personalized education[4][5]. While the application of MAS in education is growing, a comprehensive, synthesized view of its specific use in learning assessment is conspicuously absent. Existing work often lacks a critical analysis of the field's architectural evolution and the persistent challenges impeding its adoption[6].

This paper fills this critical gap. Through a systematic literature review, we construct a clear evolutionary map of MAS for assessment, from early rule-based systems to modern ecosystems integrated with Large Language Models (LLMs). We introduce a taxonomy that classifies these systems not just by function, but by critically analyzing the success factors and inherent limitations of representative systems at each stage. This analysis culminates in a unified synthesis of unresolved challenges and a forward-looking research agenda aimed at building more effective, equitable, and explainable assessment systems.

2 Survey Methodology

This systematic survey aims to answer key questions regarding the architectural evolution, functional applications, underlying technologies, and unresolved challenges of MAS in educational assessment. We conducted a literature search between January 2010 and June 2025 across IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and arXiv. The core search query combined terms like ("multi-agent system" OR "intelligent agent") with ("learning assessment" OR "formative assessment") and ("education" OR "e-learning"). Papers were included if they explicitly detailed the design, application, or evaluation of an MAS as a central component for learning assessment. We excluded theses, editorials, non-peer-reviewed works, and papers where MAS was used only for peripheral administrative tasks. This rigorous process yielded a final corpus of key papers forming the basis for our thematic analysis (See Table 1).

3 Thematic Analysis: Evolution of MAS in Assessment Applications

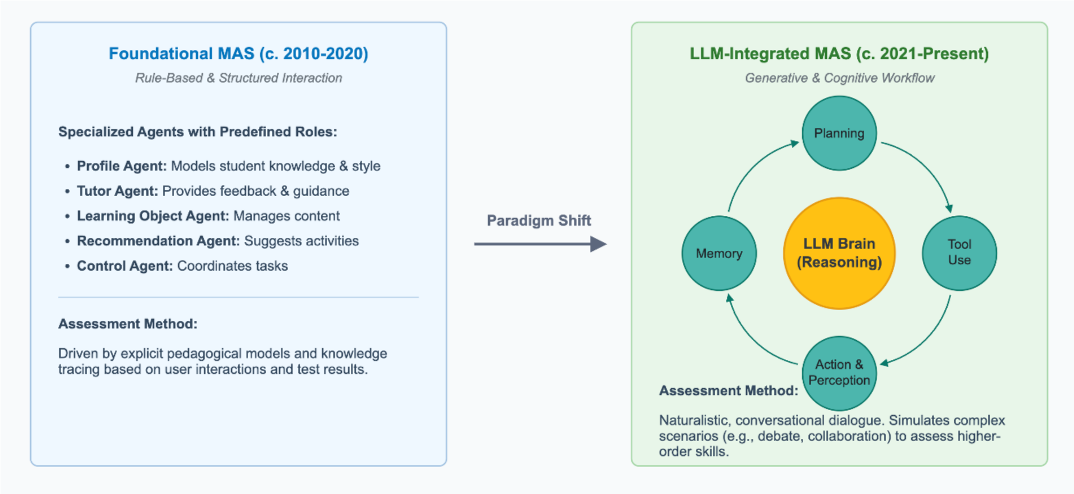
Our analysis reveals a clear evolutionary trajectory from structured, rule-driven systems to dynamic, generative ecosystems powered by LLMs (Figure 1). We structure this evolution by assessment function, critically examining representative systems from each era to identify their core strengths and weaknesses.

EKI 2025, Chongqing, China

ACM ISBN 979-8-4007-1584-6/2025/07
https://doi.org/10.1145/3765325.3765390

Table 1: Inclusion and Exclusion Criteria.

Criteria	Inclusion	Exclusion
Relevance	Explicitly focus on the design, application, or evaluation of a Multi-Agent System as a central component for learning assessment.	MAS is mentioned only peripherally or used for non-assessment tasks like administrative scheduling or simple content delivery without an assessment loop.
Publication	Peer-reviewed journal articles, conference papers, or highly-cited preprints.	Theses, dissertations, editorials, and non-peer-reviewed workshop papers.
Time Frame	Published between January 2010 and June 2025 to capture both foundational work and the most recent advancements.	Published before 2010.
Language	Published in English.	Not available in English.


Foundational MAS (c. 2010-2020)
Rule-Based & Structured Interaction
Specialized Agents with Predefined Roles:
Assessment Method:
Paradigm Shift
LLM-Integrated MAS (c. 2021-Present)
Generative & Cognitive Workflow
Planning
Tool Use
Action & Perception
Memory
LLM Brain (Reasoning)
Assessment Method:

- **Profile Agent:** Models student knowledge & style
- **Tutor Agent:** Provides feedback & guidance
- **Learning Object Agent:** Manages content
- **Recommendation Agent:** Suggests activities
- **Control Agent:** Coordinates tasks

Driven by explicit pedagogical models and knowledge tracing based on user interactions and test results.

Naturalistic, conversational dialogue. Simulates complex scenarios (e.g., debate, collaboration) to assess higher-order skills.

Figure 1: The architectural evolution of Multi-Agent Systems for learning assessment.

3.1 Personalized Formative Assessment: From Rule-Based Tutors to Conversational Mentors

This remains the most prevalent application, focusing on providing ongoing, real-time feedback to guide the learning process.

Foundational Era (c. 2010–2020): Early systems were characterized by their reliance on predefined rules and explicit pedagogical models, featuring specialized agents with hard-coded roles. A typical architecture, as exemplified in systems like [7][8][9], included a Profile Agent to model student knowledge based on interaction history, a Tutor Agent to execute teaching strategies and deliver templated feedback, and a Control Agent to coordinate tasks [4][5].

- Key Success Factor: This modular, rule-based design offered predictability and pedagogical consistency. The separation of concerns allowed for clear, interpretable logic.
- Limitation: Rigidity. These systems struggled with unanticipated student queries and complex reasoning. Adapting them to new domains required significant manual authoring of rules and content, posing a major scalability bottleneck.

LLM-Integrated Era (c. 2021–Present): The integration of LLMs marks a paradigm shift towards generative and cognitive workflows. Systems like SRLAgent [10][11] leverage a powerful LLM-based agent to enhance self-regulated learning skills through gamified, conversational interaction. The agent facilitates goal-setting, strategy execution, and self-reflection in a naturalistic dialogue.

- Key Success Factor: Conversational flexibility and contextual awareness. These agents can handle ambiguity and generate nuanced, human-like formative feedback that adapts in real-time, moving beyond rigid scripts.
- Limitation: Inherited risks from LLMs. The potential for generating factually incorrect or pedagogically unsound feedback (hallucination) is a critical barrier to trust [10]. Furthermore, their "black box" reasoning process makes it difficult for educators to understand or verify the basis of an assessment decision [12][13][14][15].

3.2 Collaborative Learning Assessment: Simulating Team Dynamics

This emerging domain was largely infeasible in the foundational era but has been unlocked by modern LLM-based systems.

LLM-Integrated Era (c. 2021–Present): MAS can now instantiate multiple agents with distinct personas to simulate group work

Table 2: Unresolved Technical And Ethical Challenges.

Challenge Category	Description of Challenge	Representative Literature
Safety & Trust	**Hallucination and Misinformation**: LLM-agents can generate plausible but incorrect feedback, a critical barrier to trust. The maintainability of code produced by LLMs is also a concern [21].	[10], [19]
Bias & Equity	**Algorithmic Bias**: Agents trained on biased data may perpetuate or amplify societal biases in assessment, unfairly penalizing certain demographic groups.	[10], [13]
Privacy	**Sensitive Data Handling**: These systems process vast amounts of sensitive student data. Robust privacy-preserving mechanisms are essential but complex to implement.	[10]
Interpretability	**The "Black Box" Problem**: The reasoning processes of complex LLM-agents are often opaque, making it difficult for educators to understand *why* an agent made a assessment decision.	[13]
Scalability	**Computational Cost and Complexity**: Deploying sophisticated, multi-agent LLM systems for thousands of students concurrently presents significant computational and financial challenges [22].	[14]

and assess collaboration skills. For instance, systems can simulate multi-agent code reviews [16] or student debates [17]. In the code review scenario, agents playing roles like 'Senior Developer' and 'QA Tester' interact to generate review comments and code revisions, assessing a student's ability to engage in constructive technical dialogue.

- Key Success Factor: The ability to simulate complex, dynamic social interactions at scale, providing a testbed for skills that are difficult to evaluate with traditional methods.
- Limitation: The validity gap. While conceptually compelling, there is scarce empirical evidence that agent-agent dynamics accurately reflect or can be used to validly assess human-human collaboration. Closing this gap is a crucial research imperative.

3.3 Summative Assessment & Academic Integrity: The Distributed Proctor

A niche but critical application is the use of MAS for automated proctoring of high-stakes online exams.

Hybrid Approach: Systems like the one described in [18] typically employ a hybrid MAS. They use multiple, specialized agents collaborating to ensure exam integrity. For example, one agent may handle identity verification via facial recognition, another monitors gaze tracking, and a third agent analyzes network traffic. These agents communicate to flag a confluence of suspicious events for a human proctor's review.

- Key Success Factor: Efficiency and real-time processing. Distributing the monitoring tasks across multiple agents allows for more robust and scalable real-time anomaly detection than a monolithic system.
- Limitation: Significant ethical and privacy challenges. These systems can be highly intrusive and are susceptible to algorithmic bias, potentially misinterpreting neurodivergent behaviors or cultural differences as cheating, thereby threatening assessment equity.

4 Future Research Agenda and Conclusion

Our analysis reveals a field in transition, moving from systems of delegation (where agents execute predefined tasks) to systems of cognition (where agents reason and collaborate) [19]. While this evolution unlocks powerful new assessment capabilities, it also surfaces significant unresolved challenges that must be addressed to ensure responsible deployment. These challenges, summarized in Table 2, directly inform our proposed research agenda.

These challenges demand a concerted research effort focused on building systems that are not only intelligent but also transparent, equitable, and trustworthy. We propose the following forward-looking research agenda:

1. Explainable and Trustworthy AI (XAI) for Assessment: The risks of hallucination and the "black box" problem directly threaten user trust. Future work must focus on designing agent architectures that generate human-understandable justifications for their feedback and assessment decisions. This includes developing frameworks that allow educators to audit, understand, and even fine-tune an agent's assessment criteria[20][21][22].
2. Advanced Retrieval and Knowledge Integration: To combat hallucination and reliance on potentially outdated internal knowledge, integrating Retrieval-Augmented Generation (RAG) is critical. Research is needed to build comprehensive educational knowledge bases and develop effective fusion strategies that combine retrieved, verifiable information with an LLM's generative capabilities to produce more robust and contextually relevant outputs [24][25].
3. Hybrid and Universal Agent Architectures: To balance the reliability of rule-based systems with the flexibility of LLM-agents, hybrid architectures are a promising path. Research should explore effective patterns for combining rule-based agents (for core knowledge and deterministic tasks) with LLM-agents (for nuanced dialogue). This approach could also mitigate the high computational costs associated with pure LLM systems.

4. Ethical Frameworks and Governance-by-Design: Ethical considerations like bias and privacy cannot be afterthoughts. Research must focus on proactively embedding technical mechanisms for bias detection, data minimization, and privacy-preserving computation into the system's design. This must be complemented by developing clear governance models for educational institutions deploying these systems [23][24][25].
5. Domain-Specific and Longitudinal Studies: To move beyond conceptual validation, large-scale, longitudinal deployments in authentic educational settings are essential. We need studies that measure the long-term impact of continuous MAS-based formative assessment on student outcomes and validate simulated collaborative assessments against real human group dynamics [11][26].

5 CONCLUSION

In conclusion, MAS represents a significant paradigm shift for educational assessment. The future, however, lies not in replacing human educators but in augmenting their capabilities. By automating the intensive, data-driven work of continuous assessment, these systems can empower teachers to focus on mentorship, inspiration, and the uniquely human aspects of student support. The research agenda outlined here provides a concrete roadmap toward realizing this vision.

References

[1] F. Kamalov, D. S. Calonge, L. Smail, D. Azizov, D. R. Thadani, T. Kwong, and A. Atif, "Evolution of AI in Education: Agentic Workflows," *arXiv preprint arXiv: 2504.20082*, 2025.

[2] T. Šalamon, *Design of Agent-Based Models.* Repin, Czech Republic: Bruckner Publishing, 2011.

[3] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-Agent Systems: A Survey," *IEEE Access*, vol. 6, pp. 28573–28593, 2018. doi: 10.1109/ACCESS.2018.2831228.

[4] M. Caridi and A. Sianesi, "Multi-agent systems in production planning and control: An application to the scheduling of mixed-model assembly lines," *International Journal of Production Economics*, vol. 68, no. 1, pp. 29–42, 2000. doi: 10.1016/S0925-5273(99)00097-3.

[5] V. Maldonado, J. Cedeño, F. Gamarra, E. Moncayo, and G. Grijalva, "Multi-Agent Systems: A Survey About Its Components, Frameworks, and Methodologies," *IEEE Access*, vol. 12, pp. 58495–58515, 2024. doi: 10.1109/ACCESS.2024.3396865.

[6] A. Roychoudhury, C. S. Păsăreanu, M. Pradel, and B. Ray, "AI Software Engineer: Programming with Trust," *arXiv preprint arXiv:2502.13767*, 2025.

[7] S. Prompiengchai, C. Narreddy, and S. Joordens, "A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI," *arXiv preprint arXiv: 2505.23405*, 2025.

[8] N. Ouherrou, A. El Haddadi, and H. Cha, "Towards multi-agent system for learning object recommendation," *Procedia Computer Science*, vol. 238, pp. 644–651, 2024. doi: 10.1016/j.procs.2024.04.110.

[9] M. Al-Emran and K. Shaalan, "Enhancement of online education system by using a multi-agent approach," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100042, 2022. doi: 10.1016/j.caeai.2022.00012.

[10] Z. Chu *et al.*, "LLM Agents for Education: Advances and Applications," *arXiv preprint arXiv:2503.11733*, 2025.

[11] W. Ge *et al.*, "SRLAgent: Enhancing Self-Regulated Learning Skills through Gamification and LLM Assistance," *arXiv preprint arXiv:2506.09968*, 2025.

[12] A. Yehudai *et al.*, "Survey on Evaluation of LLM-based Agents," *arXiv preprint arXiv:2503.16416*, 2025.

[13] G. Liang and Q. Tong, "LLM-Powered AI Agent Systems and Their Applications in Industry," *arXiv preprint arXiv:2505.16120*, 2025.

[14] Y. Yang, Q. Peng, J. Wang, and W. Zhang, "Multi-LLM-Agent Systems: Techniques and Business Perspectives," *arXiv preprint arXiv:2411.14033*, 2024.

[15] P. Zhao, Z. Jin, and N. Cheng, "An In-depth Survey of Large Language Model-based Artificial Intelligence Agents," *arXiv preprint arXiv:2309.14365*, 2023.

[16] Z. Yang, C. Gao, Z. Guo, Z. Li, K. Liu, X. Xia, and Y. Zhou, "A Survey on Modern Code Review: Progresses, Challenges and Opportunities," *arXiv preprint arXiv: 2405.18216*, 2024.

[17] J. Du, G. Xu, W. Liu, D. Zhou, and F. Liu, "Enhancing Online Learning Through Multi-Agent Debates for CS University Students," *Applied Sciences*, vol. 15, no. 11, p. 5877, 2025. doi: 10.3390/app15115877.

[18] I. Khabbachi, "An Intelligent Solution based on a Multi-agent System for the Proctoring of Online Exams," in *Proc. 7th Int. Conf. Smart Digital Environ. (ICSDE)*, Tetouan, Morocco, May 2023.

[19] L. Yan *et al.*, "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review," *arXiv preprint arXiv:2303.13379*, 2023.

[20] F. Härer, "Specification and Evaluation of Multi-Agent LLM Systems – Prototype and Cybersecurity Applications," *arXiv preprint arXiv:2506.10467*, 2025.

[21] K. Shivashankar and A. Martini, "Better Python Programming for all: With the focus on Maintainability," *arXiv preprint arXiv:2408.09134*, 2024.

[22] A. Tillmann, "Literature Review Of Multi-Agent Debate For Problem-Solving," *arXiv preprint arXiv:2506.00066*, 2025.

[23] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–111, 2024. doi: 10.1111/bjet.13370.

[24] Z. Yang, S. Chen, C. Gao, Z. Li, X. Hu, K. Liu, and X. Xia, "An Empirical Study of Retrieval-Augmented Code Generation: Challenges and Opportunities," *ACM Transactions on Software Engineering and Methodology*, 2025. doi: 10.1145/3717061.

[25] D. Quinn *et al.*, "Accelerating Retrieval-Augmented Generation," *arXiv preprint arXiv:2412.15246*, 2024.

[26] S. Hu, T. Huang, F. İlhan, S. F. Tekin, G. Liu, R. R. Kompella, and L. Liu, "A Survey on Large Language Model-Based Game Agents," *arXiv preprint arXiv:2404.02039*, 2024.