IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

A Survey on AI-Based Answer Sheet Evaluation and Result Management Systems

Diya Kubal[1], Vaidehi Kokare[2], Aayush Dhotre[3], Dr. Riyazahemed Jamadar[4]
AISSMS Institute of Information Technology, Pune, India
kubaldiya6@gmail.com, kokarevaidehi2@gmail.com
dhotreaayush123@gmail.com, riyaz.jamadar@aissmsioit.org

Abstract: *Artificial Intelligence (AI) in teaching can transform traditional methods of assessment, particularly the methods of evaluation of the descriptive and open-ended responses of the students. Large language models (LLMs) can read, interpret and generate text in a human-like way, which means that automated grading systems can be both more scalable and objective than in the past when it was performed by human evaluators. Beyond accuracy, transparent, reliable and capable of producing actionable feedback post in accordance with pedagogical standards, many LLMs offer a black box design, which can actively raise concerns about fairness, interpretability and transparency, and the cost/benefit balance of processing speed and analytical depth can limit learners and institutions in their use.*

An emerging solution, which is the focus of a review in the present paper, is dual-process LLM frameworks of automated assessment. These hybrid systems are a combination of a bigger, more complex system that would perform more comprehensive rubric-based assessment and a system that would make detailed feedback with a slightly small model which could be used to score first assessments fast. To provide researchers and educators with a comprehensive outline of the development of reliable and efficient AI-powered assessment devices.

Keywords: Automated Grading, Answer Sheet Evaluation, Natural Language Processing, Large Language Models, Dual-Process Framework, Education Technology

I. INTRODUCTION

Due to the recent scramble of the past decade on the fast-growing educational technological developments, traditional learning and assessment paradigm has been altered. The most revolutionary tools among these technological advances are the large language models (LLMs) and natural language processing (NLP). Tasks, which were previously labor intensive and subjective, can now be fully automated through reinforcements of the unmatched ability of the LLaMA to comprehend, summarize, and produce human-like text by using LLMs such as BERT, GPT, and LLaMA. One of the hardest applications of these models in classroom relates to assessment of descriptive and open-ended student responses that require understanding the context, logical integrity, and nuances in student expression and makes manual evaluation of responses tedious, unreliable, and prone to the influence of human bias, unlike multiple-choice or objective answers [12].The use of AI in assessment promises significant benefits, including the ability to be scaled, objective, and efficient. Automation can handle a considerable number of submissions by students within a relatively short period of time and provide fairly reliable grading of various responses. However, multiple challenges of the application of the LLM in assessing education exist. The black-box condition of LLMs also puts transparency, interpretability, and trust in question, as educators need to understand why a particular score was given to a particular piece of information [13]. And there is too a fundamental trade-off between speed and accuracy that small models are quicker, or at least partially correct answers, whereas large models with high capacity provide high-quality and comprehensive evaluations, at an enormous computational cost and slow speed. In order to get past these limitations, a viable strategy would involve application of dual process or hybrid frameworks. In such systems the initial scoring is

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

done by a lightweight, fast model that is able to handle most of the responses in a short period of time. Unclear or not sure responses are forwarded to a larger, more powerful model that can perform a detailed rubric based analysis and feedback. This two-level solution adopts all the good things of the single-model solutions and also adopts the best things of the single-model solutions. The stated survey is supposed to present an in-depth analysis of dual-process LLM structures of automated descriptive answer evaluation. We syntactically analyse the available literature, classify evaluation systems based on scoring methodology and architecture, and discuss the merits and demerits of dual-model systems. We also talk about real-world implementation strategies, problems, and open questions, such as fairness, data bias, interpretability, and how to add them to learning management systems. Lastly, we talk about new trends like interactive AI feedback, multimodal assessment. This paper consolidates existing knowledge, functioning as a practical guide for researchers, educators, and practitioners seeking to create reliable, transparent, and efficient AI-powered assessment systems suitable for large-scale implementation. To solve this problem, dual-process or hybrid frameworks have come up. This paper examines a system that utilizes two separate AI models:

1) Fast Model: A BERT-based network that has been fine-tuned for quick, high-throughput initial scoring.

2) Slow Model: A carefully adjusted LLaMA model that only runs when needed for in-depth, rubric-based analysis and feedback generation.

We give a practical guide for making strong automated evaluation systems by focusing on this dual-model approach.

II. OBJECTIVES

The primary aim of this survey and the suggested system is as follows:

A. To automatize and properly evaluate answers with LLMs:

The primary aim of this survey and the suggested system is as follows. To automatize and properly evaluate answers with LLMs. The system makes use of Large Language models (LLMs) and more complex NLP methods to compare student answers to model semantically, identify answers, identify mistakes and impose relevant grades. Through context recognition, paraphrasing can be identified and recreating minute details, the system reduces human subjectivity, makes large datasets consistent, and saves time and effort that would have been spent on manual grading.

B. To implement a dual-module scoring system:

To implement a dual-module scoring system. There is a Fast Module that is built in architecture for initial scoring and a Slowness Module to go deep into the analysis where ambiguous or poorly confident responses are processed. This dual-process design is efficient and analytical depth, which enables high confidence responses to be processed swiftly as well as with care that those that are complicated, and half right, or unorthodox solutions are critically reviewed.

C. To be able to give feedback and personalized insights:

To be able to give feedback and personalized insights. The system produces other than scores student individual feedback drawing on strong, weak points, conceptual gaps and errors. It can suggest specific improvement, reading materials, or practice activity, which encourages independent learning and supporting teachers to personalize teaching plans.

D. To offer interactive dashboards and analytics:

To offer interactive dashboards and analytics. Student and teacher dashboards provide extensive information, visual performance tracking, result management, analytics. The teachers are able to break down trends in classes.

III. SCOPE OF THE SURVEY

The dual-process Large Language Model (LLM) frameworks for the automated assessment of descriptive and open-ended student responses in educational contexts are the main topic of this survey. The main goal is to investigate systems that combine a slow, larger model for in-depth rubric-based evaluation and personalized feedback with a quick, lightweight model for quick preliminary scoring. By focusing on this hybrid architecture, the survey highlights the useful advantages, difficulties, and factors to be taken into account when implementing dual-process frameworks in actual classroom settings and standardized assessment scenarios. The scope covers a number of automated grading

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

features, such as the calculation of semantic similarity, confidence-based answer escalation, and integration with interactive whiteboards for teachers and students. In order to strike a balance between speed and grading accuracy, the survey looks at how slow models handle ambiguous cases selectively while fast models handle most answers efficiently [5]. It also considers domain-specific adjustments of the LLCs, including fine-tuning to specific subjects, grade levels, or academic domains, in addition to the preprocessing techniques required to put the student responses into a normalized state and increase the models' performance. It is also the scope that determines the current weaknesses and advancements in the field. Though the key area of this review is a text based descriptive response, it acknowledges that multimodal assessment that involves equations, diagrams, handwritten responses and audio submissions is increasingly gaining popularity as a potential route of research. Also, the survey addresses such considerations as transparency, fairness, and interpretability, which are essential aspects in educational AI. Reliable usage and widespread implementation of AI-generated scores rely on the process of ensuring that the scores can be clarified to teachers and students. The teaching usefulness of the specific feedback provided by dual-process frameworks is also evaluated; this will help students to detect the gaps in their concepts, to build their critical thinking skills, and to obtain essential recommendations on how to improve [7]. To facilitate a holistic educational ecosystem, the survey also includes the integration with Learning Management Systems (LMS) that can provide real-time analytics, performance tracking, adaptive learning suggestions. Also, this survey will help bridge the gap between the theoretic and the real world of application of LLM within the educational settings. It provides information on best practice to develop scalable AI-powered grading systems; based on providing an analysis of existing architecture, workflow architecture and performance measures. It also determines the problems of data bias, model interpretability, and its correspondence with existing pedagogical systems to guide future research and development activities. Besides automated grading, the scope emphasizes the enhancement of the learning process as well as more learner-centered approach.

TABLE I: COMPARISON OF EXISTING APPROACHES AND PROPOSED FRAMEWORK

Author / Paper	Main Idea	Advantage	Disadvantage / Gap
Early NLP-Based Grading	Keyword matching and TF-IDF scoring for automated grading	Computationally efficient; foundation for automated grading	Limited contextual understanding; cannot process paraphrased answers
Deep Learning-Based Evaluation	RNNs, LSTMs, and early transformer models for contextual response processing	Captures contextual dependencies	High computational complexity; lacks architectural efficiency
Proposed Framework	Two-process LLM design integrating fast similarity scoring and rubric-based grading	Balances efficiency and accuracy; scalable and interpretable feedback generation	Focused on two-process design; exploration of hybrid grading space left for future work

IV. SCOPE OF RESEARCH

The recent decades were characterized by the growth of interest regarding automated evaluation of student responses, which can be viewed as a sign of the current evolution of assessment methods and educational technology. Most of the early evaluation systems relied on the use of statistical and rule-based methods, including TF-IDF (Term Frequency-Inverse Document Frequency) scoring, matching on keywords and bag-of-words methods. These techniques provided the first effort at automating the grading process by comparing student responses and reference answers and identifying exact matches or overlaps. Even though such systems were simple to adopt and computationally inexpensive, they did not have a lot of pedagogical significance as they could not understand the semantic content of the responses or the raw score [11]. Despite these limitations, these methods precondition the future researches on automated educational evaluation with the accent on the fact that the number of works to be graded manually can be reduced; more significantly, the efficiency of academic evaluation may be enhanced. Machine learning and deep learning practices

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, March 2026 Impact Factor: 8.2

began to transform the sphere of automated evaluation dramatically. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks allowed the capacity to process word sequences and determine contextual dependencies in responses of students. The architectures were more successful in comparison with earlier rule-based approaches, allowing the systems to evaluate grammar, coherence, logical structure, and general relevance. These models were still hard pressed to capture long-range dependencies and minute nuances of semantics, however, in complex, descriptive and even open-ended responses. One of the major developments was the introduction of transformer-based architecture such as BERT, GPT and LLaMA. With pre-trained embeddings and self-attention, these models could generate rich contextual representations of text, which made semantic comparisons between response of the students and reference solutions more accurate. Transformers demonstrated high ability in identifying errors, partial feedbacks, and generation of justifications of complex answers besides scoring and grading assignments. Smaller models, while faster, frequently compromised accuracy and depth of analysis, while large models' high computational cost and slower processing time limited their applicability in large-scale or real-time assessment scenarios [9]. There are still a number of issues in spite of the advancements made by transformer-based and deep learning techniques. Since educators need to have faith in the logic underlying automated scores, interpretability and explainability of the model are important considerations. Adoption in formal educational settings may be hampered by the fact that many high-performing models function as "black boxes," offering scores without explicit explanations [3]. Furthermore, there is still a trade-off between speed and accuracy: small models are effective but might miss subtle errors or partial correctness, while large models are accurate but computationally costly. In order to balance these trade-offs, hybrid or dual-process frameworks have thus been the focus of recent research. In these systems, most answer evaluations are handled by a quick, lightweight model, while ambiguous or low-confidence responses are subjected to a thorough rubric-based analysis by a slower, more potent model. Besides being more productive, this architecture provides more credible and open feedback, which is what modern teachers need. On balance, the literature demonstrates a clear evolution of simple rule-based systems into more complex system models of transformers and indicates the achievements, as well as persisting limitations of automated evaluation. The dual-process LLM framework, which extends this development, offers a practical, scalable, and interpretable solution that can be seen to have the benefits of the previous models without having as many disadvantages [10].

TABLE II: DUAL-PROCESS LLM FRAMEWORK OVERVIEW

Module	Purpose / Features	Advantages
Fast Model	Fast scoring through semantic embeddings to calculate similarity between student and reference answers	Efficient processing; processes high volumes; initial grading
Slow Model	Rubric-based scoring through a large language model for unclear answers	High accuracy; in-depth reasoning; personalized feedback generation

V. RELATED WORK

In the last 2 decades, much research has been conducted on the application of artificial intelligence (AI) in educational assessment, and it indicates a steady increase in computational capabilities, practices, and procedures. Most of the early techniques involved statistical and keyword-matching techniques, including cosine similarity, TF-IDF scoring, and bag-of-words models. These systems could be used to perform a simple automated evaluation by comparing student responses with reference responses and hence, it had several disadvantages. Interestingly, they have often produced strict and superficial scoring [11], were not semantically aware, and could not process paraphrased or contextually sensitive answers. The more advanced approaches were introduced with deep learning. Transformer-based models, including BERT and GPT, recurrent neural networks (RNNs), and long short-term memory networks (LSTMs), allow learning contextual embeddings and semantic representations of student responses. These models allowed making a more advanced assessment of grammar, coherence, the relevance of the content and logical structure possible [2]. Recent surveys have indicated the effectiveness of these architectures in generating partial feedback as well as providing stable scoring. Such models did have new challenges as well. High capacity models are not suitable to handle

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

large batches of responses because they are slow and expensive to run. Smaller models however are more efficient at the expense of depth and accuracy and highlights the trade-off between speed and accuracy that is persisting. Another important problem identified in the literature is explainability. The educators require proper evaluation approaches that reinforce grades and highlight specific strengths or weaknesses in the answers of students. Most deep learning methods are black-box methods, even though they are very precise; this does not promote their adoption and confidence in real education [4]. Existing reviews do a good job of defining the problem space, but they frequently fall short of offering workable architectural solutions that strike a balance between interpretability, speed, and accuracy. This has spurred research into dual-process or hybrid frameworks, which combine slower, more potent models for in-depth analysis and feedback with quick, lightweight models for bulk scoring. By thoroughly examining dual-process LLM frameworks and evaluating their advantages, disadvantages, and potential for scalable implementation in educational assessment, our survey fills this knowledge gap.

VI. DUAL-PROCESS LLM FRAMEWORKS IN PRACTICE

A. The Fast Model: Speed and Efficiency

The majority of student responses can be handled by the Fast Model with little computational expense. Usually, it makes use of a lightweight neural network, like a distilled BERT variant (all-MiniLM-L6-v2), which converts the reference answer and the student answer into high-dimensional vector embeddings. A preliminary score and a confidence metric are then produced by computing semantic similarity metrics, such as cosine similarity. This module's primary strength is its high throughput, which enables it to effectively evaluate hundreds or thousands of responses in real time. The workload of the more expensive computationally Slow Model is also reduced in that clear-cut cases, those answers which are either very similar or which are clearly incorrect are handled rapidly. This design allows scalability and allows automated evaluation of standardized tests and large classrooms. The Fast Model can also be adapted on topic-specific datasets to make it more accurate to particular subjects or topics[1]. An example of such an embedding can be altered to find formulaic expressions or domain-specific terms within science or mathematics assessment. This allows initial scoring with more accuracy without significantly increasing the time to compute. At the same time, the Fast Model can be regularly updated with new responses of students, which improves its operation over time through incremental learning an essential aspect of dynamic learning settings.

Fig. 1 Workflow of System

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

B. The Slow Model: Depth and Explainability

Response assessment is directed to the Slow Model when Fast Model scores the responses with low-confidence scores. This is a module built on a larger transformer-based LLM, such as LLaMa, that has been modified directly to educational evaluation. Compared to the Fast Model, the Slow Model is a comprehensive rubric-based analysis of all the aspects of the student response in consideration of the predetermined standards. It generates customized, useful feedback by highlighting certain errors, absent ideas or false memories. It is also selective in its use since it only applied on complex or ambiguous cases, yet computationally more expensive. Also, another necessary component is the fairness and transparency of educators, which is achieved through the Slow Model [6]. By providing explanations about the assignment scores, teachers will be able to understand the reasoning behind AI score and override results in case of necessity. Moreover, it may give students an opportunity to get improvement options or hints, allowing it to create a more personal and interactive learning experience. Also, multi criteria scoring, including grammar, factual correctness, and logical flow that is often ignored by single model automated grading, is facilitated by this module.

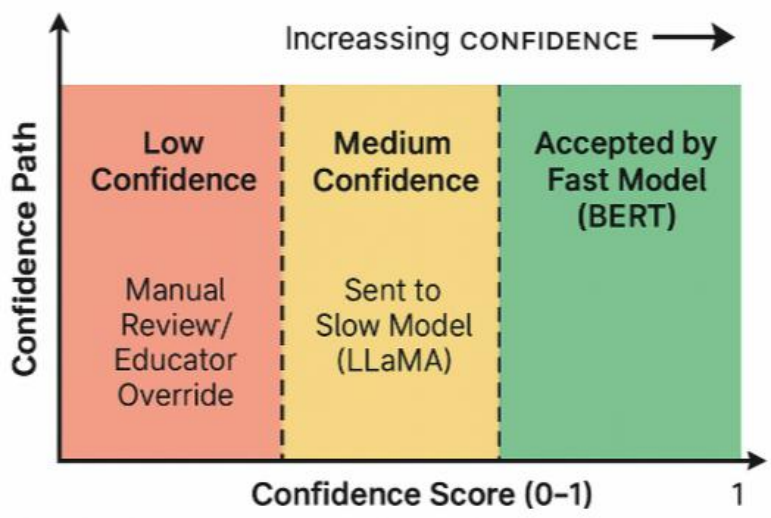

Fig. 2 Confidence-based Evaluation Flow

C. Integration and Workflow

The dual-process system works according to the confidence. Each student answer is processed in the Fast Model first that evaluates its confidence and scores quickly. Medium cases get automatically placed to Slow Model and those with high or low similarity are considered final. This integration ensures the optimal trade-off between accuracy and efficiency by removing unnecessary computation and maintaining grading fidelity. The workflow can also be enhanced by dynamic confidence thresholds that vary based on the preferences of teachers, trends in the performance of the classes or the complexity of the subject matter. The system can also maintain a historical database of graded responses, which can be readily updated in order to allow trend analysis and also uncover some of the common misunderstandings with student groups [8]. The integration with learning management systems (LMS) allows teachers to monitor student progress, trends, and make focused interventions with real-time dashboards. Moreover, students can freely access their detailed feedback and motivate self-assessment and progressive learning. It can be also enhanced in the future to add multimodal assessment capabilities, the support of diagrams, handwritten text and verbal responses, and it would become a versatile instrument of modern learning settings.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

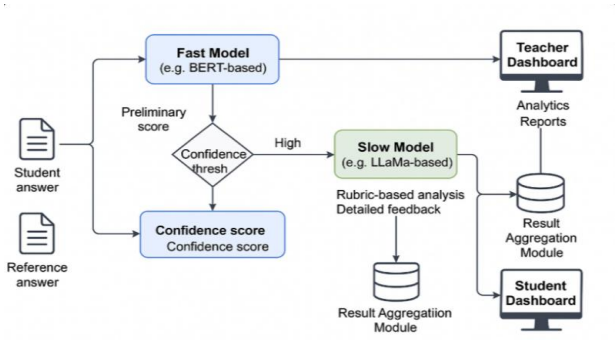

Fig. 3 System Architecture

VII. CHALLENGES AND OPEN PROBLEMS

• Grading Faithfulness: It can be ensured that the scores and justifications provided by automated grading systems are not misleading the logic of model predictions. misalignment may result in inaccurate feedback which will compromise the credibility of the system.

• Data Quality and Bias: The quality of the training data is an important determinant of the effectiveness of AI grading. The incomplete, noisy and biased datasets may lead to unfair punishment of students or misunderstanding of real answers.

• Academic Integrity and Standardization: It is hard to set fair standards of assessment in various subjects, education boards and types of questions. The absence of standardization might lead to inconsistent grading.

• Educator Trust: Teachers would not be eager to use automated grading systems, in case they cannot comprehend or justify the logic behind the scores they are given. The issues of transparency and smooth integration of the LMS play a vital role in building trust.

• Handling Open-Ended Responses: It is common to find AI models being unable to evaluate well creative writing, free-form thought, or unorthodox approaches to solving problems. Systems need to be flexible and on the right track.

• Scalability and Adaptability: Educational centers produce huge numbers of student feedback. The systems should be able to scale up effectively to meet the changing curricula, assessment patterns and the types of questions.

• Error Detection and Correction: Auto grading can sometimes be inaccurate concerning responses. Reliability should be ensured through providing the option of manual review, incorporation of feedback loop and correction of scores.

VIII. FUTURE WORK

• Interactive Comments: Students can be able to converse with the systems, offer clarifications, pointers, or detailed instructions to help them better understand it, which is more than merely obtaining a fixed score.

• Multimodal Evaluation: In order to offer a more comprehensive evaluation of the skills of the students, grading capacities must be broadened to handwritten answers, diagrams, audio records, or video files.

• Privacy-Preserving AI: Federated learning, differential privacy, or encrypted computation can be used to train the models without revealing any personal information about students.

• Adaptive Learning Integration: Increasing the effectiveness of learning process through the insights of grading to determine what each student has weak areas and propose learning resources as well as custom-tailored courses.

• Cross-Language and Cultural Flexibility: Developing models that accurately rate responses in diverse languages, dialects and culture to access AI evaluation everywhere in the world.

• Explainable AI Improvements: Establishing trusts between students and teachers through clarifying the reasons behind the grades and feedback of each item.

• Continuous Model Improvement: The grading models would continue being refined with the input of the teachers and students which is in real-time, ensuring accuracy, equity and relevance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, March 2026
Impact Factor: 8.2

IX. CONCLUSION

The use of AI in education requires both transparency, efficiency, fairness, and interpretability as well as high accuracy in grading. This paper has looked at dual-process LLM models, a slow, more analytical one at the in-depth analysis and a fast model that ensures the speed of assessment. The proposed architecture preserves pedagogical importance, and it allows an effective AI-supported evaluation due to a balance between the speed and levels of reasoning. Integrating automated community with explanatory feedback can allow the system to provide teachers with valuable insights on student performance, helping to pinpoint the existing learning gaps and tailor the interventions. Moreover, such aspects as quality of data, decreasing bias and aligning the curriculum will play a significant role in making sure that a great variety of students are graded fairly, and unfair cases are avoided, and trust is developed between educators and learners. With interactive feedback, multimodal evaluation and privacy-preservation approaches integrated to AI-based assessment systems, the future of education may be revolutionized. These can process open-answer questions, cater to various learning models and even provide personalised, on-the-fly counselling to complement traditional teaching methods. Grading remains contextually relevant, accurate and reliable through feedback loops that refine the model over time. In the context of explainability, scalability and ethical compliance dual-process LLM Frameworks indicate a way forward for ethical AI application in Education. Ultimately, these systems hope to enable teachers and students by decreasing workloads and increasing efficiency while creating more transparent, inclusive, and data-informed learning environments.

.

REFERENCES

[1]. Zhang, S., Wang, X., Zhang, W., Li, C., Song, J., Li, T., Qiu, L., Cao, X., Cai, X., Yao, W., Zhang, W., Wang, X., & Wen, Y. (2025). Leveraging dual process theory in language agent framework for real-time simultaneous human-AI collaboration. arXiv preprint arXiv:2502.11882.

[2]. Gao, Y. (2025, July 15). AI and auto-grading in higher education: Capabilities, ethics, and the evolving role of educators. ASC Office of Distance Education, The Ohio State University. auto-grading-higher-capabilities-ethics-and-evolving-role-educators

[3]. Garzón, J. (2025). Systematic review of artificial intelligence in education. MDPI.

[4]. Wang, S. (2024). Artificial intelligence in education: A systematic literature review. Computers in Human Behavior, 132, 107254.

[5]. Gao, Y. (2023). Artificial intelligence and the future of teaching and learning: Insights and recommendations. U.S. Department of Education, Office of Educational Technology.

[6]. Melo-López, V. A. (2025). The impact of artificial intelligence on inclusive education. MDPI.

[7]. Vieriu, A. M. (2025). The impact of artificial intelligence on students' learning processes and academic performance. MDPI.

[8]. Merino-Campos, C. (2025). The impact of artificial intelligence on personalized learning in higher education. MDPI.

[9]. Amofa, B. (2025). Navigating the complexity of generative artificial intelligence in teaching and learning. MDPI.

[10]. Bellini-Leite, S. C. (2024). Dual process theory for large language models: An exploration of cognitive frameworks. SAGE Open, 14(1), 21582440231120604.

[11]. Panayides, A. S., et al. (2020). AI in medical imaging informatics: Current challenges and future directions. IEEE Journal of Biomedical and Health Informatics, 24(7), 1837–1857.

[12]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NIPS) (pp. 4765–4774).

[13]. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. (2024). arXiv preprint arXiv:2401.06431.