ISSN 2308-5916. Mathematical and computer modelling.
Series: Technical sciences. 2026. Issue 29. P. 62-70.

UDC 004.89:378.147
DOI: 10.32626/2308-5916.2026-29.62-70

Kolodii R. I.

ORCID: 0009-0002-4628-6422,
PhD student, Lviv Polytechnic National University, Lviv, Ukraine,
E-mail: roman.i.kolodii@lpnu.ua

Vykliuk Y. I.

ORCID: 0000-0003-4766-4659,
D. Sc., Professor, Lviv Polytechnic National University, Lviv, Ukraine,
E-mail: yaroslav.i.vykliuk@lpnu.ua

METHOD OF ITERATIVE MULTI-AGENT VERIFICATION OF GRADES IN VIRTUAL LEARNING ENVIRONMENTS BASED ON EXPLAINABLE ARTIFICIAL INTELLIGENCE

The paper develops a method of iterative multi-agent verification of grades that ensures transparency and reliability of automated open-response assessment in virtual learning environments (VLE). The relevance of the problem of opaque decision-making by large language models (LLMs) and their tendency to generate factually incorrect statements in educational assessment tasks is substantiated. A formal model of the VLE assessment subsystem as a multi-agent system comprising three specialized agents (evaluator agent, verifier agent, and explainer agent) is proposed. For each agent, input-to-output mapping functions are defined. The MultiAgentGrading algorithm implementing a four-phase assessment procedure is developed: initial generation using chain-of-thought strategy (ante-hoc component), critical analysis by the verifier (post-hoc component), iterative refinement, and pedagogical aggregation of the result. The method combines built-in and post-hoc explainability mechanisms in a unified agent interaction cycle, enabling minimization of hallucination risks and enhancement of assessment reproducibility. Convergence conditions for the iterative process and a safeguard mechanism against infinite loops are defined. The transition from a linear «more explanations means more trust» paradigm to a calibrated trust concept is justified, where user confidence aligns with the model's actual capabilities.

Key words: *multi-agent system, explainable artificial intelligence, large language models, virtual learning environment, automated assessment, verification, chain-of-thought, calibrated trust.*

Стаття надійшла до редакції: 18.03.2026
Рекомендовано до друку: 28.03.2026
Оприлюднено (online): 15.05.2026

Introduction.

Problem statement. The integration of artificial intelligence (AI) agents into virtual learning environments (VLE) significantly expands the possibilities for automating pedagogical processes. Among the most promising directions is automated open-response assessment using large language models (LLMs), which enables not only quantitative grading but also the generation of detailed justifications [1, 2]. However, the practical deployment of LLMs faces critical barriers: the opacity of decision-making mechanisms (the «black box» effect) and the tendency to generate factually incorrect statements known as «hallucinations» [3, 4]. In the educational context, these limitations are fundamental, as they directly affect the objectivity of assessment and the level of stakeholder trust in the system.

Analysis of recent research and publications. The issue of automated assessment using LLMs is actively investigated in current scientific literature. Works [1, 2] demonstrate that LLMs can achieve significant agreement with human experts in essay and open-response scoring. At the same time, studies [3] emphasize the need for additional verification mechanisms to detect and minimize hallucinations. In the context of explainable AI (XAI), significant contributions were made by Ribeiro et al. [5], who proposed the LIME method for local classifier interpretation, and Lundberg and Lee [6], who developed the game-theoretic SHAP approach for feature importance determination. However, these post-hoc methods are characterized by instability risks and vulnerability to manipulation [7]. The Chain-of-Thought method proposed by Wei et al. [8] demonstrated significant improvement in LLM reasoning quality through the generation of intermediate logical steps. Multi-agent approaches to educational system organization are considered in works [9, 10], which show the advantages of distributed architecture for adaptive learning.

Identification of previously unsolved parts of the general problem. Despite significant progress in each of the outlined directions, a systematic model that would simultaneously provide multi-agent coordination of the assessment process, pedagogically relevant explainability, and LLM error minimization mechanisms has not yet been presented in the literature. Existing XAI approaches are primarily oriented toward classical machine learning models and do not account for the specifics of generative language models in the educational context. Furthermore, there is a lack of formalized methods for building calibrated trust, where user confidence aligns with actual model capabilities through verification procedures [11].

The purpose of the article is to develop and theoretically substantiate a method of iterative multi-agent verification of grades in VLE that ensures transparency and reliability of automated assessment results based on the composition of built-in and post-hoc explainable AI mechanisms.

Main Material.

1. Formal model of the multi-agent knowledge assessment environment. To implement the tasks of intellectualizing the assessment process, it is proposed to consider the VLE assessment subsystem as a multi-agent system (MAS), where the final result is formed through the composition of individual agent functions [9]. This approach allows formalizing the assessment process not as a single model call, but as an iterative process of interaction between specialized entities.

Formally, the model of the intellectualized assessment environment can be represented as a tuple:

$$M_{VLE} = \langle S, A, K, Q, R, \Phi \rangle, \quad (1)$$

where $S = \{s_1, s_2, \ldots, s_n\}$ is the finite set of assessment process states; $A = \{a_1, a_2, \ldots, a_m\}$ is the set of AI agents performing specialized roles; K is the knowledge base and assessment criteria; Q is the set of control questions; R is the set of possible student responses; Φ is the set of inter-agent interaction protocols.

Agent specification. The set of agents A is divided into subsets according to their roles:

$$A = \{ a_{eval}, a_{ver}, a_{expl} \}, \quad (2)$$

where each agent implements a specific mapping function from input data to intermediate or final results.

Evaluator agent (a_eval) performs the primary response analysis function:

$$f_{eval} : Q \times R \times K \to G_{pre} \times E_{raw}, \quad (3)$$

where $G\_pre \in [0, 100]$ is the preliminary quantitative grade, and E_raw is the draft justification (initial explanation). At this stage, the Chain-of-Thought strategy [8] is applied for generating logical conclusions.

Verifier agent (a_ver) is responsible for detecting factual errors in the draft justification:

$$f_{ver} : Q \times R \times E_{raw} \to \{0,1\} \times C_{feedback}, \quad (4)$$

where {0, 1} is the binary validity status, and $C\_feedback$ is the corrective comment (feedback for the evaluator agent).

Explainer agent (a_expl) aggregates verified data and forms the final explanation:

$$f_{expl} : G_{pre} \times E_{raw} \times C_{feedback} \to G_{final} \times E_{final}, \quad (5)$$

where E_final is the verified, pedagogically adapted explanation that complies with human-centered design principles [11].

Formalization of the assessment process. The process of obtaining a grade with explanation is iterative. Let $q_i \in Q$ be the current question, and $r_i \in R$ be the student's response. Then the intellectualized assessment function is defined as the composition of agent functions:

$$F_{XAI}(q_i, r_i) = f_{expl} \circ (f_{ver} \circ f_{eval})^k (q_i, r_i, K), \quad (6)$$

where k is the number of reconciliation iterations between the evaluator and verifier.

The task reduces to minimizing the explanation error function, which depends on the discrepancy between the generated explanation and facts from the knowledge base:

$$E^* = arg\,min_E\, L(E \mid r_i, K), \quad (7)$$

where minimization is achieved through iterative agent interaction until the stopping criterion is met: absence of critical verifier remarks or exhaustion of the iteration limit.

2. Method of iterative multi-agent grade verification. Based on the developed formal model, an automated assessment method is proposed, the key feature of which is the procedural separation of grade generation and validation stages. The method implements a hybrid approach based on the composition of two explainability mechanisms [5-7]:

- *built-in justification generation (ante-hoc component)* – generation of structured decision-making logic directly during the assessment process;
- *post-hoc verification (post-hoc component)* – independent verification of the generated explanation by a separate agent to detect errors or contradictions.

Agent interaction procedure. Let the system input be the tuple (q, r, K), where q is the control question, r is the student's response, K is the assessment criteria. The procedure is executed in four stages.

Stage 1 (Drafting Phase). The evaluator agent receives input data and generates a draft grade *O_draft*. The Chain-of-Thought strategy [8] is applied, requiring the agent to explicate intermediate reasoning before forming a score. This implements *ante-hoc* explainability, where the decision logic is formed simultaneously with the decision itself.

Stage 2 (Verification Phase). The draft *O_draft* is passed to the verifier agent, which operates in a «blind review» mode. Its task is to verify the correspondence of facts in the explanation to the student's response content and criteria. If a contradiction is detected, the status is set to *Invalid*.

Stage 3 (Refinement Loop). In case of a negative verification result, the feedback mechanism is activated. The evaluator agent receives struc-

tured comments from the verifier and generates a corrected version. The cycle repeats until stopping conditions are met: obtaining *Valid* status (agent consensus) or reaching the iteration limit *N_max*.

Stage 4 (Finalization Phase). The explainer agent transforms the verified conclusion into a pedagogically correct form containing the grade, justification, and recommendations for knowledge improvement.

3. Algorithmic implementation of the method. For a formal description of the method's logic, we present the pseudocode of the MultiAgentGrading algorithm.

```
    Algorithm 1. MultiAgentGrading
    Input: q – question; r – response; K – criteria;
N_max – iteration limit
    Output: FinalOutput – {grade, explanation, recom-
mendations}

    1. iter_count := 0; status := INVALID; feedback
:= NULL
    2. draft := a_eval.generate_draft(q, r, K)
    3. WHILE (status = INVALID) AND (iter_count <
N_max) DO:
    4. result := a_ver.audit(draft, r, K)
    5. IF result.is_valid = TRUE THEN:
    6. status := VALID; BREAK
    7. ELSE:
    8. feedback := result.comments
    9. draft := a_eval.refine(draft, feedback)
    10. iter_count := iter_count + 1
    11. END WHILE
    12. IF status = INVALID THEN:
    13. draft.add_warning("Requires manual review")
    14. FinalOutput := a_expl.format_response(draft)
    15. RETURN FinalOutput
```

The algorithm implements an iterative scheme where each iteration verifies the consistency between the initial assessment and criteria. The iteration count limitation by the ***N_max*** parameter prevents infinite loops in cases where agents fail to reach consensus. When the iteration limit is exhausted, the system automatically adds a warning about the need for manual review, ensuring safe system behavior in edge cases.

The proposed approach ensures synergy of built-in and post-hoc mechanisms within multi-agent interaction, aimed at increasing assessment reproducibility, minimizing hallucination risks [3, 4], and generating explanations adapted to human-centered assessment principles [11, 12].

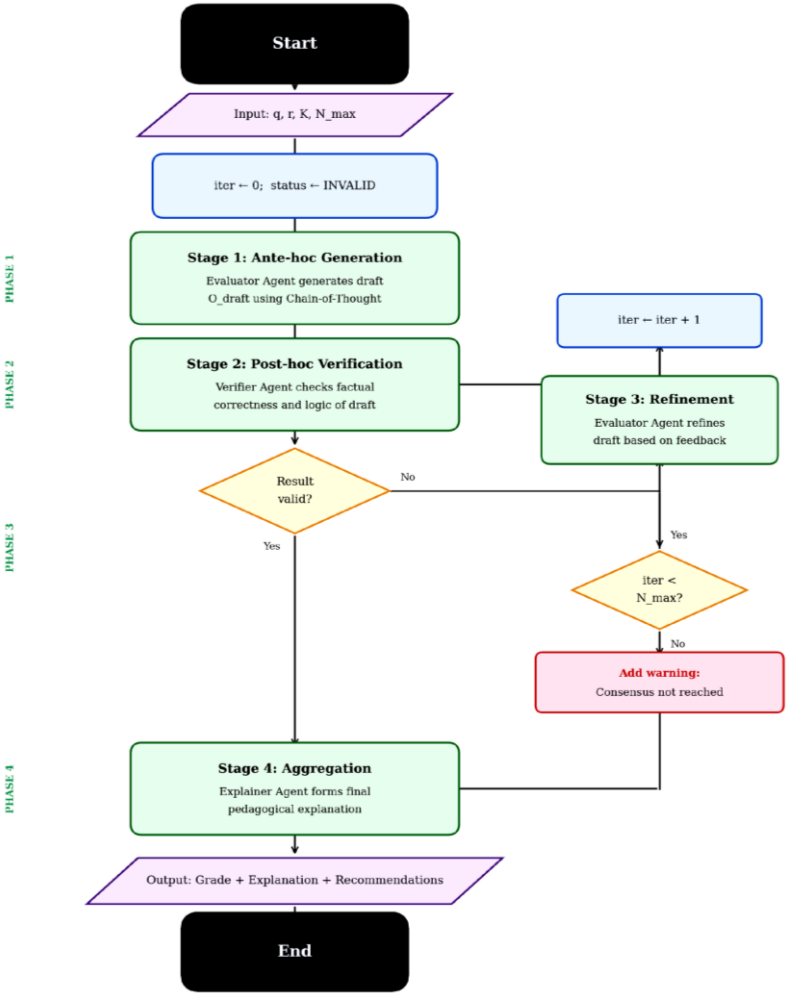

Fig. 1. *Flowchart of the multi-agent grade verification algorithm*

Table 1

Comparison of assessment approaches with explanation

Characteristic	**Single LLM call**	**Post-hoc XAI (LIME/SHAP)**	**Proposed method**
Decision transparency	Low	Medium	High
Hallucination mitigation	None	Partial	Systematic

Continuation of the table 1

Explanation stability	Variable	Unstable (sensitive to perturbations)	Controlled
Calibrated trust	Not provided	Not provided	Built-in mechanism
Computational cost	Low	Medium	Moderate (2-4 LLM calls)

The comparative analysis (Table 1) demonstrates that the proposed method combines the advantages of built-in and post-hoc explainability, providing systematic hallucination control at moderate computational costs.

Conclusions. The paper develops a method of iterative multi-agent grade verification aimed at enhancing the transparency and reliability of automated assessment in virtual learning environments. The main results of the work are as follows.

1. A formal model of the VLE assessment subsystem as a multi-agent system $M\_VLE = \langle S, A, K, Q, R, \Phi \rangle$ comprising three specialized agents with defined mapping functions is proposed. The model ensures clear separation of responsibility between grade generation, verification, and pedagogical adaptation stages.
2. The MultiAgentGrading algorithm implementing a hybrid approach to explainable assessment through the composition of ante-hoc (chain-of-thought) and post-hoc (critical analysis by verifier) mechanisms is developed. The algorithm includes a safeguard mechanism against infinite loops and automatic warning about the need for manual review.
3. The transition from a linear «more explanations means more trust» paradigm to a calibrated trust concept is justified, ensuring the alignment of user confidence with the model's actual capabilities.

Prospects for further research include empirical verification of the proposed method in a real educational environment, integration with learning management platforms (particularly Moodle), and expansion of the architecture with additional specialized agents for subject-oriented assessment.

References:

1. Liew P. Y., Tan I. K. T. On Automated Essay Grading using Large Language Models. *CSAI '24: Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*. New York: Association for Computing Machinery, 2025. P. 204-211. DOI: 10.1145/3709026.3709030.
2. Pack A., Barrett A., Escalante J. Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Computers and Education: Artificial Intelligence*. 2024. Vol. 6. Art. 100234. DOI: 10.1016/j.caeai.2024.100234.

3. Huang L., Yu W., Ma W. et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*. 2025. Vol. 43. № 2. Art. 42. P. 1-55. DOI: 10.1145/3703155.
4. Pesaranghader A., Li E. Hallucination Detection and Mitigation in Large Language Models. *arXiv*. 2026. arXiv: 2601.09929. URL: https://arxiv.org/-abs/2601.09929
5. Ribeiro M. T., Singh S., Guestrin C. «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. P. 1135-1144. DOI: 10.1145/2939672.2939778.
6. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 2017. Vol. 30. P. 4768-4777.
7. Samek W., Müller K.-R. Towards Explainable Artificial Intelligence. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, 2019. (Lecture Notes in Computer Science; Vol. 11700). P. 5-22. DOI: 10.1007/978-3-030-28954-6_1.
8. Wei J., Wang X., Schuurmans D. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*. 2022. Vol. 35. P. 24824-24837.
9. Zhang R. Multi-Agent Systems for Learning Assessment in Education: A Comprehensive Survey. *EKI '25: Proceedings of the 2025 3rd International Conference on Educational Knowledge and Informatization*. 2025. P. 388-391. DOI: 10.1145/3765325.3765390.
10. López-Goyez J. P., González-Briones A., Demazeau Y. An Adaptive Multi-Agent Architecture with Reinforcement Learning and Generative AI for Intelligent Tutoring Systems: A Moodle-Based Case Study. *Applied Sciences*. 2026. Vol. 16. № 3. Art. 1323. DOI: 10.3390/app16031323.
11. Alfredo R., Echeverria V., Jin Y. et al. Human-Centred Learning Analytics and AI in Education: A Systematic Literature Review. *Computers and Education: Artificial Intelligence*. 2024. Vol. 6. Art. 100215. DOI: 10.1016/j.caeai.2024.100215.
12. Wang X., Wei J., Schuurmans D. et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *The Eleventh International Conference on Learning Representations (ICLR 2023)*. 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw

МЕТОД ІТЕРАТИВНОЇ МУЛЬТИАГЕНТНОЇ ВЕРИФІКАЦІЇ ОЦІНОК У ВІРТУАЛЬНИХ НАВЧАЛЬНИХ СЕРЕДОВИЩАХ НА ОСНОВІ ПОЯСНЮВАНОГО ШТУЧНОГО ІНТЕЛЕКТУ

У статті розроблено метод ітеративної мультиагентної верифікації оцінок, що забезпечує прозорість та надійність автоматизованого оцінювання відкритих відповідей у віртуальних навчальних середовищах (ВНС). Обгрунтовано актуальність проблеми непрозорості рішень великих мовних моделей (ВММ) та їх схильності до генерації фактоло-

гічно некоректних тверджень у задачах освітнього оцінювання. Запропоновано формальну модель підсистеми оцінювання ВНС як мультиагентної системи, що включає три спеціалізовані агенти: агент-оцінювач, агент-верифікатор та агент-пояснювач. Для кожного агента визначено функції відображення вхідних даних у проміжні або кінцеві результати. Розроблено алгоритм MultiAgentGrading, який реалізує чотирифазну процедуру оцінювання: первинна генерація з використанням стратегії ланцюжка думок (ante-hoc компонент), критичний аналіз верифікатором (post-hoc компонент), ітеративна корекція та педагогічна агрегація результату. Метод поєднує вбудовані та постфактум-механізми пояснюваності в єдиному циклі взаємодії агентів, що дозволяє мінімізувати ризик галюцинацій та підвищити відтворюваність оцінювання. Визначено умови збіжності ітеративного процесу та запобіжний механізм проти зацикленню. Обґрунтовано перехід від лінійної парадигми пояснень до концепції каліброваної довіри, за якої рівень впевненості користувача узгоджується з реальними можливостями моделі.

Ключові слова: *мультиагентна система, пояснюваний штучний інтелект, великі мовні моделі, віртуальне навчальне середовище, автоматизоване оцінювання, верифікація, ланцюжок думок, каліброваність довіри.*